Light color acclimation: a key process in the global ocean distribution of *Synechococcus* cyanobacteria

Théophile Grébert[a], Hugo Doré[a], Frédéric Partensky[a], Gregory K. Farrant[a,1], Emmanuel S. Boss[b], Marc Picheral[c], Lionel Guidi[c], Stéphane Pesant[d,e], David J. Scanlan[f], Patrick Wincker[g], Silvia G. Acinas[h], David M. Kehoe[i] and Laurence Garczarek[a,2]

[a]Sorbonne Universités-Université Paris 06 & Centre National de la Recherche Scientifique (CNRS), UMR 7144, Marine Phototrophic Prokaryotes Team, Station Biologique, CS 90074, 29688 Roscoff cedex, France; [b]Maine In-situ Sound and Color Lab, University of Maine, Orono, ME 04469; [c]Sorbonne Universités-Université Paris 06 & CNRS, UMR 7093, Observatoire océanologique, 06230 Villefranche-sur-mer, France; [d]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany; [e]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany; [f]University of Warwick, School of Life Sciences, Coventry CV4 7AL, UK; [g]Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), Institut de Génomique, Genoscope, 91057 Evry, France; [h]Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM), Consejo Superior de Investigaciones Científicas (CSIC), Barcelona ES-08003, Spain; [i]Department of Biology, Indiana University, Bloomington, IN 47405.

[1]Present address: Food Safety, Environment, and Genetics, Matís Ltd., 113 Reykjavík, Iceland.

[2]To whom correspondence should be addressed. Email: laurence.garczarek@sb-roscoff.fr.

Classification: Biological Sciences/Environmental Sciences

Keywords: marine cyanobacteria, metagenomics, light quality, phycobilisome, *Tara* Oceans

Short title: Biogeography of *Synechococcus* pigment types

Abstract

Marine *Synechococcus* cyanobacteria are major contributors to global oceanic primary production and exhibit a unique diversity of photosynthetic pigments, allowing them to exploit a wide range of light niches. However, the relationship between pigment content and niche partitioning has remained largely undetermined due to the lack of a single-genetic marker resolving all pigment types (PTs). Here, we developed and employed a novel and robust method based on three distinct marker genes (*cpcBA*, *mpeBA* and *mpeW*) to estimate the relative abundance of all known *Synechococcus* PTs from metagenomes. Analysis of the *Tara* Oceans dataset allowed us, for the first time, to reveal the global distribution of *Synechococcus* PTs and to define their environmental niches. Green-light specialists (PT 3a) dominated in warm, green equatorial waters, whereas blue-light specialists (PT 3c) were particularly abundant in oligotrophic areas. Type IV chromatic acclimaters (CA4-A/B), which are able to dynamically modify their light absorption properties to maximally absorb green or blue light, were unexpectedly the most abundant PT in our dataset and predominated at depth and high latitudes. We also identified populations in which CA4 might be nonfunctional due to the lack of specific CA4 genes, notably in warm high-nutrient low-chlorophyll areas. Major ecotypes within clades I-IV and CRD1 were preferentially associated with a particular PT, while others exhibited a wide range of PTs. Altogether, this study provides important insights into the ecology of *Synechococcus* and highlights the complex interactions between vertical phylogeny, pigmentation and environmental parameters that shape *Synechococcus* community structure and evolution.

Significance Statement

Understanding the functional diversity of specific microbial groups at the global scale is critical yet poorly developed. By combining the considerable knowledge accumulated through recent years on the molecular bases of photosynthetic pigment diversity in marine *Synechococcus*, a major phytoplanktonic organism, with the wealth of metagenomic data provided by the *Tara* Oceans

expedition, we have been able to reliably quantify all known pigment types along its transect and provide the first global distribution map. Unexpectedly, cells able to dynamically change their pigment content to match the ambient light color were ubiquitous and predominated in many environments. Altogether, our results unveiled the role of adaptation to light quality on niche partitioning in a key primary producer.

\body

Introduction

Marine *Synechococcus* is the second most abundant phytoplankton group in the world's oceans and constitutes a major contributor to global primary production and carbon cycling (1, 2). This genus displays a wide genetic diversity and several studies have shown that among the ~20 clades defined based on various genetic markers, five (clades I-IV and CRD1) predominate *in situ* and can be broadly associated with distinct sets of physico-chemical parameters (3–5). In a recent study, we further defined Ecologically Significant Taxonomic Units (ESTUs), i.e. organisms belonging to the same clade and co-occurring in the field, and highlighted that the three main parameters affecting the *in situ* distribution of these ESTUs were temperature and availability of iron and phosphorus (6). Yet, marine *Synechococcus* also display a wide pigment diversity, suggesting that light could also influence their ecological distribution, both qualitatively and quantitatively (7, 8).

This pigment diversity comes from differences in the composition of their main light-harvesting antennae, called phycobilisomes (PBS; 7–9). These water-soluble macromolecular complexes consist of a central core anchoring at least six radiating rods made of several distinct phycobiliproteins, i.e. proteins to which specific enzymes (phycobilin lyases) covalently attach chromophores called phycobilins (7, 10). Although the PBS core is conserved in all marine *Synechococcus*, rods have a very variable composition, and three main pigment types (PTs) are usually distinguished (Fig. S1; 7, 11). In PT 1, PBS rods are solely made of phycocyanin (PC, encoded by the *cpcBA* operon) and bear the red-light absorbing phycocyanobilin (PCB; A_{max} = 620 nm) as the sole chromophore. In PT 2, rods are made of PC and phycoerythrin I (PE-I, encoded by *cpeBA*) and attach both PCB and the green-light absorbing phycoerythrobilin (PEB; A_{max} = 550 nm). All other marine *Synechococcus* belong to PT 3 and have rods made of PC, PE-I and PE-II (encoded by *mpeBA*) that bind PCB, PEB and the blue-light absorbing phycourobilin (PUB; A_{max} = 495 nm; Fig. S1). Several subtypes can be defined within PT 3,

based on the fluorescence excitation ratio at 495 nm and 545 nm (hereafter $Ex_{495:545}$; Fig. S1), a proxy for the PUB:PEB ratio. This ratio is low ($Ex_{495:545} < 0.6$) in subtype 3a (green light specialists), intermediate in subtype 3b ($0.6 \leq Ex_{495:545} < 1.6$) and high ($Ex_{495:545} \geq 1.6$) in subtype 3c (blue light specialists; 7, 11). Additionally, strains of subtype 3d are able to change their PUB:PEB ratio depending on ambient light color, a process called type IV chromatic acclimation (hereafter CA4), allowing them to maximally absorb blue or green light (11–14). Comparative genomic analyses showed that genes involved in the synthesis and regulation of PBS rods are gathered into a dedicated genomic region, the content and organization of which correspond to the different PTs (7). Similarly, chromatic acclimation has been correlated with the presence of a small specific genomic island (CA4 genomic island) that exists in two distinct configurations (CA4-A and -B; 11). Both contain two regulators (*fciA* and *fciB*) and a phycobilin lyase (*mpeZ* in CA4-A or *mpeW* in CA4-B), thus defining two distinct CA4 genotypes: 3dA and 3dB (11, 14, 15). Finally, some strains possess a complete or partial CA4 genomic island but are not able to perform CA4, displaying a fixed $Ex_{495:545}$ corresponding to 3a, 3b or 3c phenotypes (11).

As there is no correspondence between pigmentation and core genome phylogeny (7, 16, 17), deciphering the relative abundance and niche partitioning of *Synechococcus* PTs in the environment requires specific approaches. In the past 30 years, studies have been based either on i) proxies of the PUB:PEB ratio as assessed by flow cytometry (18–20), fluorescence excitation spectra (21–27), epifluorescence microscopy (28), or ii) phylogenetic analyses of *cpcBA* or *cpeBA* (17, 29–34). These studies showed that PT 1 is restricted to and dominates in low salinity surface waters and/or estuaries, which are characterized by a high turbidity resulting in a red wavelengths-dominated light field (18, 22, 31–38), whereas PT 2 is found in coastal shelf waters or in the transition zones between brackish and oceanic environments with intermediate optical properties (18, 27, 34, 36–39). Finally, PT 3 with increasing PUB:PEB ratio are found over gradients from onshore mesotrophic waters, characterized by green light dominance, to offshore oligotrophic waters, where blue light penetrates the deepest (19–24, 28, 36, 38, 40). Some authors reported an increase in the PUB:PEB ratio with

105 depth (19, 21, 24), while others observed a constant ratio throughout the water column, a variability
106 potentially linked to the location, water column features and/or environmental parameters (22, 25,
107 28).

108 However, these analyses based on optical properties could only describe the distribution of
109 high- and low-PUB populations without being able to differentiate green (3a) or blue light (3c)
110 specialists from CA4 cells (3d) acclimated to green or blue light, while genetic analysis solely based on
111 *cpcBA* and/or *cpeBA* could not differentiate all PTs. For instance, only two studies have reported CA4
112 populations *in situ* either in the western English Channel (17) or in sub-polar waters of the western
113 Pacific Ocean (29) but none of them were able to differentiate CA4-B from high PUB (i.e. 3c)
114 populations. As a consequence, the global relative abundance of the different *Synechococcus* PTs,
115 particularly CA4, and the link between genetic and pigment diversity have remained largely unclear.

116 Here, we analyzed 109 metagenomic samples collected from all major oceanic basins during the
117 2.5-yr *Tara* Oceans (2009-2011) expedition (41) using a bioinformatic pipeline combining a
118 metagenomic read recruitment approach (6, 42) to recruit single reads from multiple PBS gene
119 markers and placement of these reads in reference trees to assign them to a given PT. This pipeline
120 allowed the first description of the worldwide distribution of all known *Synechococcus* PTs, as well as
121 of their realized environmental niches (*sensu* 43). This study provides a synoptic view of how a major
122 photosynthetic organism adapts to natural light color gradients in the ocean.

Results

A novel, robust approach for estimating pigment types abundance from metagenomes

We developed a multi-marker approach combining phylogenetic information retrieved from three different genes or operons (*cpcBA*, *mpeBA* and *mpeW*; Fig. 1 and Datasets 1-2) to overcome the issue of fully resolving the whole range of PTs. While *cpcBA* discriminated PT 1, 2 and 3 (Fig. 1A), only the *mpeBA* operon, a PT 3 specific marker, was able to distinguish the different PT 3 subtypes (Fig. 1B), though as for *cpeBA* it could not differentiate PT 3dB (CA4-B) from PT 3c (i.e. blue light specialists; 11, 29). The *mpeW* marker was thus selected to specifically target PT 3dB and, by subtraction, enumerate PT 3c (Fig. 1C). Using the *cpcBA* marker, members of PT 2 were split into two well-defined clusters, 2A and 2B (Fig. 1A), the latter corresponding to a purely environmental PT identified from assembled metagenomes of the Baltic Sea (38). Strains KORDI-100 and CC9616 also clustered apart from other strains in the *mpeBA* phylogeny, suggesting that they have a divergent evolutionary history from other PT 3 members (Fig. 1B). This is supported by the diverged gene content and order of their PBS rod genomic region and these strains were recently referred to as PT 3f, even though they have a similar phenotype as PT 3c ($Ex_{495:545}$ ratio ≥ 1.6; 30). To investigate the phylogenetic resolution of small fragments of these three markers, sequences were removed one at a time from the reference database, and simulated reads (150 bp long as compared to 164 bp in average for *Tara* Oceans cleaned/merged reads) generated from this sequence were assigned using our bioinformatic pipeline against a database comprising the remaining sequences. Inferred and known PTs were then compared. The percentage of simulated reads assigned to the correct PT was between 93.2% and 97.0% for all three markers, with less than 2.1-5.6% of reads that could not be classified and an error-rate below 2%, showing that all three markers display a sufficient resolution to reliably assign the different PTs (Fig. S2B, D and F).

To ensure that the different markers could be quantitatively compared in a real dataset, we examined the correlations between estimates of PT abundances using the different markers in the

109 metagenomes analyzed in this study. Total *cpcBA* counts were highly correlated (R^2=0.994, n=109; Fig. S3A) with total *Synechococcus* counts obtained with the *petB* gene, which was previously used to study the phylogeography of marine picocyanobacteria (6), and the correlation slope was not significantly different from 1 (slope: 1.040; Wilcoxon's paired difference test p-value=0.356). *cpcBA* is thus as good as *petB* at capturing the total population of *Synechococcus* reads. Moreover, counts of *cpcBA* reads assigned to PT 3 and total *mpeBA* counts (specific for PT 3) were also strongly correlated (R^2=0.996, n=109; Fig. S3B), and not skewed from 1 (slope of 0.991, Wilcoxon's p-value=0.607), indicating that *mpeBA* and *cpcBA* counts can be directly compared. Although no redundant information for PT 3dB is available with the three selected markers, another marker targeting 3dB (*fciAB*) was tested and produced results similar to *mpeW* (Fig. S3C). These results demonstrate that our multi-marker approach can be used to reliably and quantitatively infer the different *Synechococcus* PTs from short metagenomic reads, with PT 1, 2A, 2B abundances being assessed by *cpcBA* normalized counts, PT 3a, 3f and 3dA by *mpeBA* normalized counts, PT 3dB by *mpeW* normalized counts and PT 3c by the difference between *mpeBA* normalized counts for 3c + 3dB and *mpeW* normalized counts. We thus used this approach on the *Tara* Oceans metagenomes, generated from 109 samples collected at 65 stations located in the major oceanic basins (Fig. 2).

CA4 populations are widespread and predominate at depth and high latitudes

The latitudinal distribution of *Synechococcus* inferred from *cpcBA* counts is globally consistent with previous studies (2, 6, 44), with *Synechococcus* being present in most oceanic waters, but quasi absent (< 20 *cpcBA* counts) beyond 60°S (Southern Ocean stations TARA_082 to TARA_085; Fig. 2B). Overall, the number of recruited *cpcBA* reads per station was between 0 and 8,151 (n=63, median: 449, mean: 924, sd: 1478) for surface and 0 and 3,200 (n=46, median:170, mean: 446, sd: 664) for deep chlorophyll maximum (DCM) samples, respectively. Stations with less than 30 *cpcBA* reads were excluded from further analysis.

PT 1 and 2, being both known to be mostly found and abundant in coastal waters (29, 36, 38, 45), were expectedly almost absent from this dataset (total of 15 and 513 *cpcBA* reads, respectively; Fig. 2A-B) since the *Tara* cruise sampling was principally performed in oceanic waters. While PT 2A was mostly found at the surface at one station off Panama (TARA_141, 417 out of 6,637 reads at this station; Fig. 2B), PT 2B was virtually absent (total of 3 *cpcBA* reads) from our dataset and might thus be confined to the Baltic Sea (38). This low abundance of PT 1 and 2B precluded the correlation analysis between their distribution and physico-chemical parameters. PT 3 was by far the most abundant along the *Tara* Oceans transect, accounting for 99.1 ± 1.4% (mean ± sd) of *cpcBA* reads at stations with ≥30 *cpcBA* read counts. Interestingly, several PT 3 subtypes often co-occurred at a given station.

PT 3a (green light specialists) totaled 20.3% of read counts, with similar abundance between surface (20.5%) and DCM (19.4%) samples, and was particularly abundant in intertropical oceanic borders and regional seas, including the Red Sea, the Arabian Sea and the Panama/Gulf of Mexico area (Fig. 2B). Correlation analyses show that this PT is consistently associated with high temperatures but also with greenish (as estimated from a low blue to green downwelling irradiance ratio, $Irr_{495:545}$), particle-rich waters (high particle backscattering at 470 nm and beam attenuation coefficient at 660 nm; Fig. 3). Still, in contrast with previous studies that reported the distribution of low-PUB populations (19, 21, 23, 24, 26, 27), this PT does not seem to be restricted to coastal waters, explaining its absence of correlation with chlorophyll concentration and colored dissolved organic matter (cDOM).

Blue light specialists (PT 3c) appear to be globally widespread, with the exception of high latitude North Atlantic waters, and accounted for 33.4% of reads, with a higher relative abundance at the surface (36.8%) than at the DCM (23.3%, Fig. 2A). This PT is dominant in transparent, oligotrophic, iron-replete areas such as the Mediterranean Sea as well as South Atlantic and Indian Ocean gyres (Figs. 2B and 4C). In the South Pacific, PT 3c was also found to be predominant in the

198 Marquesas Islands area (TARA_123 and 124), where the coast proximity induced a local iron
199 enrichment (6). Consistently, PT 3c was found to be positively associated with iron concentration,
200 high temperature and DCM depth and anti-correlated with chlorophyll fluorescence, nitrogen
201 concentrations, net primary production (NPP) as well as other related optical parameters, such as
202 backscattering at 470 nm and beam attenuation coefficient at 660 nm (Fig. 3). Despite its rarity, PT 3f
203 seems to thrive in a similar environment, with the highest relative abundances in the Indian Ocean
204 and Mediterranean Sea (Figs. 2B and 4C). Its occurrence in the latter area might explain its strong
205 anti-correlation with phosphorus availability.

206 Both CA4 types, 3dA and 3dB, which represented 22.6% and 18.9% of reads respectively, were
207 unexpectedly widespread and could locally account for up to 95% of the total *Synechococcus*
208 population (Figs. 2, 4C and S4). In contrast to blue and green light specialists, both CA4 types were
209 proportionally less abundant at the surface (19.8% and 17.5%, for 3dA and 3dB, respectively) than at
210 depth (30.9% and 22.9%). Interestingly, PT 3dA and 3dB generally displayed complementary
211 distributions along the *Tara* Oceans transect (Fig. 2B). PT 3dA was predominant at high latitude in the
212 northern hemisphere as well as in other vertically mixed waters such as in the Chilean upwelling
213 (TARA_093) or in the Agulhas current (TARA_066 and 68; Fig 2B). Accordingly, PT 3dA distribution
214 seems to be driven by low temperature, high nutrient and highly productive waters (high NPP,
215 chlorophyll *a* and optical parameters), a combination of physico-chemical parameters almost
216 opposite to those observed for blue light specialists (PT 3c; Fig. 3). In contrast, PT 3dB shares a
217 number of characteristics with PT 3c, including the anti-correlation with nitrogen concentration and
218 association with iron availability (as indicated by both a positive correlation with [Fe] and negative
219 correlation with the iron limitation proxy Φsat; Fig. 3), consistent with their widespread occurrence
220 in iron replete oceanic areas. Also noteworthy, PT 3dB was one of the sole PT (with 3f) to be
221 associated with low photosynthetically available radiation (PAR).

222

Niche partitioning of *Synechococcus* populations rely on a subtle combination of ESTU and PT niches

We previously showed that temperature, iron and phosphorus availability constituted major factors influencing the diversification and niche partitioning of *Synechococcus* ESTUs (i.e. genetically related subgroups within clades that co-occur in the field; 6). Yet, these results cannot be extended to PTs since the pigment content does not follow the vertical phylogeny (7). In order to decipher the respective roles of genetic and pigment diversity in *Synechococcus* community structure, we examined the relationships between ESTUs and PTs *in situ* abundances through correlation and NMDS analyses (Fig. 4A-B) and compared their respective distributions (Figs. 4C and S4).

Interestingly, all PTs are either preferentially associated with or excluded from a subset of ESTUs. PT 2A is found at low abundance at a few stations along the *Tara* Oceans transect and, when present, it is seemingly associated with the rare ESTU 5.3B (Fig. 4A), an unusual PT/ESTU combination so far only observed in metagenomes from freshwater reservoirs (46). PT 3a is associated with ESTUs EnvBC (occurring in low iron areas) and IIA, the major ESTU in the global ocean (Fig. 4A), a result consistent with NMDS analysis, which shows that PT 3a is found in assemblages dominated by these two ESTUs (indicated by red and grey backgrounds in Fig. 4B), as well as with independent observations on cultured strains (Dataset 3). PT 3c is associated with ESTU IIIA (the dominant ESTU in P-depleted areas), as observed on many isolates (Dataset 3), and is also linked, like PT 3f, with ESTUs IIIB and WPC1A, both present at lower abundance than IIIA in P-poor waters (Fig. 4A). PT 3f is also associated with the newly described and low-abundance ESTU XXA (previously EnvC; Fig. S5; 4, 6). Both PT 3f and ESTU XXA were rare in our dataset but systematically co-occurred, in agreement with the fact that the only culture representative of the latter clade belongs to PT 3f (Dataset 3).

PT 3dA appears to be associated with all ESTUs from clades CRD1 (specific to iron-depleted areas) as well as with those representative of coastal and cold waters (IA, IVA, IVC), but is anti-correlated with most other major ESTUs (IIA, IIIA and –B, WPC1A and 5.3B; Fig. 4A). This pattern is

opposite to PT 3dB that is preferentially found associated with ESTU IIA, IIB and 5.3A, but not in CRD1A or –C (Fig. 4A). Thus, it seems that the two types of CA4 are found in distinct and complementary sets of ESTUs. Interestingly, our analysis might suggest the occurrence of additional PTs not isolated so far, since a number of reads (0.7% and 2.7% of *cpcBA* and *mpeBA* counts, respectively, Fig. 2A) could not be assigned to any known PTs. For instance, while most CRD1C seem preferentially associated with PT 3dA, a fraction of the population could only be assigned at the PT 3 level (Fig. 4A). Similarly, a number of reads could not be assigned to any known PT in stations rich in ESTU 5.3A and XXA, although one cannot exclude that this observation might be due to a low number of representative strains, and thus PT reference sequences, for these ESTUs.

The preferred association of PTs with specific ESTUs is also well illustrated by some concomitant shifts of PTs and ESTU assemblages. For instance, in the wintertime North Atlantic Ocean, the shift from 3dB-dominated stations on the western side (TARA_142 and TARA_146-149) to 3dA-dominated stations near European coasts (TARA_150 to 152) and North of Gulf stream (TARA_145) is probably related to the shift in ESTU assemblages occurring along this transect, with ESTU IIA being gradually replaced by ESTU IVA (Fig. 4C; see also 6). Similarly, the takeover of CRD1C by IIA in the Marquesas Island area (TARA_123 to 125), which is iron-enriched with regard to surrounding high-nutrient low-chlorophyll (HNLC) waters (TARA_122 and 128), perfectly matched the corresponding replacement of PT 3dA by 3c. However, in several other cases, PT shifts were not associated with a concomitant ESTU shift or vice versa. One of clearest examples of these dissociations is the transect from the Mediterranean Sea to the Indian Ocean, where the entry in the northern Red Sea through the Suez Canal triggered a sharp shift from a IIIA- to a IIA-dominated community (TARA_030 and 031), which was not accompanied by any obvious change in PTs. Conversely, a sharp rise in the relative abundance of PT 3a was observed in the southern Red Sea/northeastern Indian Ocean (TARA_033 to 038) without changes in the large dominance of ESTU IIA. Altogether, this strongly suggests that a subtle combination of ESTUs and PTs respective niche occupancy is responsible for the observed niche partitioning of *Synechococcus* populations.

Deficient chromatic acclimaters are dominant in HNLC areas

Although our results clearly indicate that CA4 cells represent a large proportion of the *Synechococcus* community in a wide range of ecological niches, this must be somewhat tempered by the fact that, in culture, about 30% of the strains possessing a CA4-A or B genomic island are not able to chromatically acclimate (Dataset 3; 11). Some of these natural mutants have an incomplete CA4 genomic island (Fig. S6K). For example, strains WH8016 (ESTU IA) and KORDI-49 (WPC1A) both lack the CA4-A specific lyase-isomerase MpeZ, an enzyme shown to bind a PUB molecule on PE-II (14), and display a green light specialist phenotype (PT 3a, $Ex_{495:545}$ ~ 0.4) whatever the ambient light color (11). However, since they possess a PT 3a *mpeBA* allele, reads from field WH8016- or KORDI-49-like cells are adequately counted as PT 3a (Fig. S6K). Another CA4-deficient strain, BIOS-E4-1 (ESTU CRD1C), possesses *mpeZ* and a 3dA *mpeBA* allele but lacks the CA4 regulators FciA and FciB as well as the putative lyase MpeY and exhibits a fixed blue light specialist phenotype (PT 3c, $Ex_{495:545}$ ~ 1.7; Fig. S6K; 11, 15). Thus, reads from such natural *Synechococcus* CA4-incapable mutants in the field are counted as 3dA using the *mpeBA* marker. Lastly, the strain MVIR-18-1 possesses a complete CA4-A island and a 3dA *mpeBA* allele but lacks *mpeU*, a gene necessary for blue light acclimation (Fig. S6K; 47). While MVIR-18-1 displays a fixed green light phenotype, reads from such *Synechococcus* are also erroneously counted as 3dA.

To assess the significance of these genotypes in the field, we compared the normalized read counts obtained for 3dA with *mpeBA*, *fciAB*, *mpeZ, mpeU* and *mpeY* (Fig. S6A-J). Overall this analysis revealed a high consistency between these different markers ($0.860<R^2<0.986$), indicating that most *mpeZ*-containing populations also contained 3dA alleles for *fciAB*, *mpeY*, *mpeU* and *mpeBA* and are therefore likely able to perform CA4. However, a number of stations, all located in HNLC areas (TARA_094, 111 and 122 to 128 in the Pacific Ocean and TARA_052 located northwest of Madagascar, Fig. 2B), displayed more than 10-fold higher *mpeBA, mpeU* and *mpeZ* counts than *fciAB*

and *mpeY* counts (Fig. S6A, B, E, F, H, I). This indicates that a large proportion or even the whole population (TARA_122 and 124) of 3dA in these HNLC areas is probably lacking the FciA/B regulators and MpeY and, like strain BIOS-E4-1 (Fig. S6K), might thus be stuck in the blue light specialist phenotype (PT 3c; 11). Conversely, station TARA_067 exhibited consistently more than twice the *fciAB* and *mpeZ* counts compared to *mpeBA*, *mpeY* or *mpeU* (Fig. S6B-E, G, H) and was a clear outlier when comparing pigment type and clade composition (Fig. S7). This suggests that the proportion of PT 3dA might have been underestimated at this station, as a significant proportion of this population probably corresponds to PT 3a genotypes that have acquired a CA4-A island by lateral gene transfer, as is seemingly the case for strains WH8016 and KORDI-49. Finally, no station exhibited markedly lower *mpeU* counts compared to all other genes, indicating that the genotype of strain MVIR-18-1 is probably rare in the oceans.

It must be noted that two out of the six sequenced CA4-B strains (WH8103 and WH8109) also have a deficient CA4 phenotype and display a constant, intermediate $Ex_{495:545}$ ratio (0.7 and 1, respectively), despite any obvious PBS- or CA4-related gene deletion (11). Accordingly, the plot of 3dB normalized read counts obtained with *mpeW* vs. *fciAB* shows no clear outlier (Fig. S3C).

Discussion

Marine *Synechococcus* display a large pigment diversity, with different PTs preferentially harvesting distinct regions of the light spectrum. Previous studies based on optical properties or on a single genetic marker could not differentiate all PTs (17, 29–31), and thus neither assess their respective realized environmental niches (43) nor the role of light quality on the relative abundance of each PT. Here, we showed that a metagenomic read recruitment approach combining three genetic markers can be used to reliably predict all major PTs. Applied to the extensive *Tara* Oceans dataset, this original approach, which avoids PCR amplification and cloning biases, allowed us to describe for the

first time the distribution of the different *Synechococcus* PTs at the global scale and to refine our understanding of their ecology.

PT 3 was found to be largely dominant over PT 1 and 2 along the oceanic *Tara* Oceans transect, in agreement with the coastal-restricted distribution of the latter PTs (18, 22, 27, 31–34, 37–39). Biogeography and correlation analyses with environmental parameters provided several novel and important insights concerning niche partitioning of PT 3 subtypes. Green (PT 3a) and blue (PT 3c) light specialists were both shown to dominate in warm areas but display clearly distinct niches, with 3a dominating in *Synechococcus*-rich stations located on oceanic borders, while 3c predominated in purely oceanic areas where the global abundance of *Synechococcus* is low. These results are in agreement with the prevailing view of an increase in the PUB:PEB ratio from green onshore mesotrophic waters to blue offshore oligotrophic waters (19–24, 26–29, 40, 48). Similarly, we showed that PT 3dB, which could not be distinguished from PT 3c in previous studies (17, 29–31), prevails in more coastal and/or mixed temperate waters than do 3c populations. The realized environmental niche of the second type of CA4 (PT 3dA) is the best defined of all PTs as it is clearly associated with nutrient-rich waters and with the coldest stations of our dataset, occurring at high latitude, at depth and/or in vertically mixed waters (e.g., TARA_068, 093 and 133). This result is consistent with a recent study demonstrating the dominance of 3dA in sub-Arctic waters of the Northwest Pacific Ocean (29), suggesting that the prevalence of 3dA at high latitude can be generalized. The decrease of PT 3c (blue light specialists) with depth is unexpected given previous reports of a constant (22, 25, 28, 49) or increasing (19, 21, 24) PUB:PEB ratio throughout the water column. However, the high abundance of CA4 can reconcile these observations with the decreased abundance of PT 3c, as cells capable of CA4 likely have a blue-light phenotype at depth. Altogether, while little was previously known about the abundance and distribution of CA4 populations in the field, here we show that they are ubiquitous, dominate in a wide range of niches, are present both in coastal and oceanic mixed waters, and overall are the most abundant *Synechococcus* PT.

348 The relationship between ESTUs and PTs shows that some ESTUs are preferentially associated
349 with only one PT, while others present a much larger pigment diversity. ESTU IIA, the most abundant
350 and ubiquitous ESTU in the field (5, 6), displays the widest PT diversity (Fig. 4B), a finding confirmed
351 by clade II isolates spanning the largest diversity of pigment content, with representative strains of
352 PT 2, 3a, 3c and 3dB within this clade (Dataset 3; see also 7, 11, 50–52). This suggests that this ESTU
353 can colonize all light color niches, an ability which might be partially responsible for its global
354 ecological success. Our current results do not support the previously observed correlation between
355 clade III and PT 3a (29) since the two ESTUs defined within this clade (IIIA and B) were associated
356 with PT 3c and/or 3f. This discrepancy could be due either to the different methods used in these
357 studies or to the occurrence of genetically distinct clade III populations in coastal areas of the
358 northwestern Pacific Ocean and along the *Tara* Oceans transect. However, the pigment phenotype of
359 strains isolated to date is more consistent with our findings (Dataset 3; 16, 36).

360 In contrast to most other PTs, the association between PT 3dA and ESTUs was found to be nearly
361 exclusive in the field, as ESTUs from clades I, IV, CRD1 and EnvA were not associated with any other
362 PT, and reciprocally PT 3dA is only associated with these clades (Fig. 4A). An interesting exception to
363 this general rule was observed in the Benguela upwelling (TARA_ 067), where the dominant ESTU IA
364 population both displays a 3a *mpeBA* allele and possesses *fciA/B* and *mpeZ* genes (Figs. S6K and S7),
365 suggesting that cells, which were initially green light specialists (PT 3a), have inherited a complete
366 CA4-A island through lateral gene transfer at this station. Interestingly, among the seven clade I
367 strains sequenced to date, three possess a 3a *mpeBA* allele, among which WH8016 also has a CA4-A
368 island but only partial (lacking *mpeZ*) and therefore not functional (11). It is thus difficult to conclude
369 whether the lateral transfer of this island, likely a rare event since it was only observed in
370 populations of the Benguela upwelling, has conferred these populations the ability to perform CA4.

371 Another important result of this study was the unsuspected importance of populations that
372 have likely lost the ability to chromatically acclimate, specifically in warm HNLC areas, which cover

wide expanses of the South Pacific Ocean (53). Interestingly, populations living in these ultra-oligotrophic environments have a different genetic basis for their consistently elevated PUB phenotype than do typical blue light specialists (i.e. PT 3c), since they have lost the CA4 regulators *fciA/B* and accumulated mutations in *mpeY,* a yet uncharacterized member of the phycobilin lyase family, as observed in strain BIOS-E4-1 (Fig. S6K; 11). This finding, consistent with the previous observation that the south Pacific is dominated by high-PUB *Synechococcus* (22), is further supported by the recent sequencing of three isolates from the Equatorial Pacific, strains MITS9504, MITS9509 (both CRD1C) and MITS9508 (CRD1A; 54), all of which contain, like BIOS-E4-1, a 3dA *mpeBA* allele, a CA4-A island lacking *fciA/B* and a partial (MITS9508) or highly degenerated (2 other MIT strains) *mpeY* gene sequence (Fig. S6K). Thus, these natural CA4-A mutants seem to have adapted to blue, ultra-oligotrophic waters by inactivating a likely energetically costly acclimation mechanism (positive selection), although we cannot exclude that it might be a consequence of the lower selection efficiency associated to the reduced effective population size of *Synechococcus* in such an extreme environment (genetic drift). If, as we hypothesize, all *Synechococcus* cells counted as 3dA at these stations are CA4-deficient, these natural mutants would represent about 15% of the total 3dA population. In contrast, CRD1-A populations of the eastern border of the Pacific Ocean (TARA_102, 109-110, 137) are likely true CA4 populations as they possess all CA4 genes (Fig. S6K).

In conclusion, our study provided novel insights into the distribution, ecology and adaptive value of all known *Synechococcus* PTs. Unexpectedly, the sum of 3dA and 3dB constituted about 40% of the total *Synechococcus* counts in the *Tara* Oceans dataset, making chromatic acclimaters (PT 3d) the most globally abundant PT, even when taking into account potential CA4-deficient natural mutants. In addition, this PT made up 95% of the *Synechococcus* population at high latitudes and was present in every one of the five major clades in the field (I, II, III, IV and CRD1). This suggests that chromatic acclimation likely confers a strong adaptive advantage compared to strains with a fixed pigmentation, particularly in vertically mixed environments and at depth at stations with a stratified water column. The occurrence of natural CA4 mutants and evidence for lateral transfer of the CA4

genomic island further support previous hypotheses that not only temperature and nutrient availability (3, 5, 6) but also light quality (7, 52) co-exert selective pressures affecting marine *Synechococcus* evolution. Thus, changes in pigment diversity could occur in response to changes in light niches by acquisition or loss of specific PBS synthesis and/or regulation genes, as previously observed for phosphorus and nitrogen transport genes in *Prochlorococcus* (55–57). Still, the complex interactions between PTs, vertical phylogeny and environmental parameters remain unclear and more work is needed to refine our understanding of the balance between the forces shaping community composition and *Synechococcus* evolution. At the boundaries of *Synechococcus* environmental niche(s), where the harshest conditions are encountered, both pigment and clade diversity are drastically reduced, and this concomitant reduction tends to support a co-selection by light quality and other environmental parameters. On the contrary, the diverse PTs occurring within some clades, as well as the co-occurrence of different PTs at most stations compared to more clear-cut clade shifts (e.g., in the Red Sea/Indian Ocean) might indicate that light quality is not the strongest selective force or that light changes are too transient to allow the dominance and fixation of a particular PT in a population. Future experimental work exploring the fitness of distinct ESTU/PT combinations under different controlled environmental conditions (including temperature, nutrients and light) might help clarifying the respective effects of these parameters on the diversification of this ecologically important photosynthetic organism.

Materials and Methods

Metagenomic samples

This study focused on 109 metagenomic samples corresponding to 65 stations from the worldwide oceans collected during the 2.5-yr *Tara* Oceans circumnavigation (2009-2011). Water sample and sequence processing are the same than in (6). Dataset 4 describes all metagenomic samples with

423 location and sequencing effort. Sequencing depths ranged from 16 × 10^6 to 258 × 10^6 reads per
424 sample after quality control and paired-reads merging, and corresponding fragments lengths
425 averaged 164 ± 20 bp (median: 168 bp).

426

427 **Databases: reference and outgroup sequences**

428 A reference database comprising the full-length gene or operon nucleotide sequences was generated
429 for each marker used in this study (*cpcBA*, *mpeBA* and *mpeW*) based on culture isolates with
430 characterized pigment type (Dataset 1). These databases comprised 83 *cpcBA* sequences (64 unique),
431 including 18 PT 1, 5 PT 2A, 19 PT 2B and 39 PT 3, 41 *mpeBA* sequences (all unique), including 11 PT
432 3a, 2 PT 3f, 11 PT 3dA and 17 PT 3dB and 5 unique *mpeW* sequences. For each marker, a reference
433 alignment was generated with MAFFT L-INS-i v6.953b (58), and a reference phylogenetic tree was
434 inferred with PhyML v. 20120412 (GTR+I+G, 10 random starting trees, best of SPR and NNI moves,
435 500 bootstraps; (59) and drawn using the ETE Toolkit (60).

436 A database of outgroups was also built, comprising paralogous sequences from marine
437 *Synechococcus* or *Prochlorococcus* as well as orthologous sequences from other marine and
438 freshwater organisms retrieved from public databases. For *cpcBA and mpeBA*, the outgroup
439 databases comprised *apcA*, *apcB*, *apcD*, *apcF and cpeBA* from marine *Synechococcus, ppeBA* from
440 *Prochlorococcus, cpcBA* and *cpeBA* from other non-picocyanobacterial organisms as well as either
441 *mpeBA* or *cpcBA* from marine *Synechococcus*, respectively (Datasets 1-2). For *mpeW*, the outgroup
442 database was made of paralogous genes (*mpeZ*, *mpeY* and *cpeY*) from marine *Synechococcus* or
443 *Prochlorococcus*, as no ortholog could be identified in public databases. Similarly, for *mpeY* and
444 *mpeZ*, the outgroup database comprised *cpeY*, *mpeW* as well as *mpeZ* or *mpeY,* respectively. The
445 outgroup database for *mpeU* comprised *cpeF* paralogous sequences from marine *Synechococcus* and
446 *Prochlorococcus.* No outgroup database was used for *fciAB*, as no paralogs or other distantly related
447 sequences were found either in marine *Synechococcus* and *Prochlorococcus* or in public databases.

Read assignation and estimation of PT abundance

Reads were preselected using BLAST+ (61) with relaxed parameters (blastn, maximum E-value of 1e-5, minimum percent identity 60%, minimum 75% of read length aligned), using reference sequences as subjects; the selection was then refined by a second BLAST+ round against databases of outgroups: reads with a best-hit to outgroup sequences were excluded from downstream analysis. Selected reads were then aligned to the marker reference alignment with MAFFT v.7.299b (--addfragments --adjustdirectionaccurately) and placed in the marker reference phylogenetic tree with *pplacer* (62). For each read, *pplacer* returns a list of possible positions (referred to as placements) at which it can be placed in the tree and their associated "likelihood weight ratio" (LWR, proxy for the probability of the placement; see *pplacer* publication and documentation for more details). Reads were then assigned to a pigment type using a custom classifier written in Python. Briefly, internal nodes of the reference tree were assigned a pigment type based on the pigmentation of descending nodes (PT of child reference sequences if the same for all of them, "unclassified" otherwise). For each read, placements were assigned to their nearest ascending or descending node based on their relative position on the edge, and the lowest common ancestor (LCA) of the set of nodes for which the cumulated LWR was greater than 0.95 (LCA of possible placements at 95% probability) was then computed. Finally, the read was assigned to the pigment type of this LCA. Different combinations of read assignment parameters (LCA at 90%, 95% or 100%; assignation of placements to the ascending, descending or nearest node) were also assessed, and resulted either in higher rates of unassigned reads or of wrongly assigned reads (Fig. S2).

Read counts were normalized by adjusted marker length: for each marker and each sequence file, counts were normalized by $(L - \ell + 1)$, with L the length of the marker gene (*cpcBA* mean length: 1053.7 bp; *mpeBA* mean length: 1054.6 bp; *mpeW* mean length: 1193.3 bp) and ℓ the mean length of reads in the sequence file. Finally, the abundance of PT 1, 2A and 2B was defined as the normalized

cpcBA read counts of these PT, the abundance of PT 3a, 3f and 3dA as the normalized *mpeBA* read counts of these PT, 3dB as the normalized *mpeW* count and 3c as the difference between the normalized *mpeBA* (3c + 3dB) read count and the PT 3dB count assessed with *mpeW*. The abundance of unclassified sequences was also taken into account. Detailed *petB* counts for clade and ESTU abundances were obtained from (6).

Read assignment simulations

For each marker, simulated reads were generated from one reference sequence at a time using a sliding window of 100, 125 or 150 bp (*Tara* Oceans mean read length: 164.2 bp; median 169 bp) and steps of 5 bp. Simulated reads were then assigned to a pigment type with the aforementioned bioinformatic pipeline, using all reference sequences except the one used to simulate reads ("leave one out" cross-validation scheme). Inferred PTs of simulated fragments were then compared to known PTs of reference sequences.

Statistical analyses

All environmental parameters used for statistical analyses are the same as in (6), except the blue to green irradiance ratio that was modeled as described in the supplementary materials and methods. Hierarchical clustering and NMDS analyses of stations were performed using R (63) packages cluster v1.14.4 (64) and MASS v7.3–29 (65), respectively. PT contingency tables were filtered by considering only stations with more than 30 *cpcBA* reads and 30 *mpeBA* reads, and only PT appearing in at least two stations and with more than 150 reads in the whole dataset. Contingency tables were normalized using Hellinger transformation that gives lower weights to rare PT. The Bray–Curtis distance was then used for ordination (isoMDS function; maxit, 100; k, 2). Correlations were

performed with R package Hmisc_3.17-4 with Benjamini & Hochberg multiple comparison adjusted p-value (66).

Acknowledgements

We warmly thank Dr. Annick Bricaud for fruitful discussions on biooptics, members of the ABiMS platform (Roscoff) for providing us an efficient storage and computing facility for our bioinformatics analyses as well as the NERC Biomolecular Analysis Facility (NBAF, Centre for Genomic Research, University of Liverpool) for sequencing some *Synechococcus* genomes used in this study. This work was supported by the French "Agence Nationale de la Recherche" Programs SAMOSA (ANR-13-ADAP-0010) and France Génomique (ANR-10-INBS-09), the French Government "Investissements d'Avenir" programs OCEANOMICS (ANR-11-BTBR-0008), the European Union's Seventh Framework Programs FP7 MicroB3 (grant agreement 287589) and MaCuMBA (grant agreement 311975), UK Natural Environment Research Council Grant NE/I00985X/1 and the Spanish Ministry of Science and Innovation grant MicroOcean PANGENOMICS (GL2011-26848/BOS). We also thank the support and commitment of the *Tara* Oceans coordinators and consortium, Agnès b. and E. Bourgois, the Veolia Environment Foundation, Région Bretagne, Lorient Agglomeration, World Courier, Illumina, the EDF Foundation, FRB, the Prince Albert II de Monaco Foundation, the *Tara* schooner and its captains and crew. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). This article is contribution number XXXX of *Tara* Oceans.

References

1. Guidi L, et al. (2016) Plankton networks driving carbon export in the oligotrophic ocean. *Nature* 532(7600):465–470.

2. Flombaum P, et al. (2013) Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc Natl Acad Sci* USA 110(24):9824–9829.

3. Zwirglmaier K, et al. (2008) Global phylogeography of marine *Synechococcus* and *Prochlorococcus* reveals a distinct partitioning of lineages among oceanic biomes. *Environ Microbiol* 10(1):147–161.

4. Mazard S, Ostrowski M, Partensky F, Scanlan DJ (2012) Multi-locus sequence analysis, taxonomic resolution and biogeography of marine *Synechococcus*. *Environ Microbiol* 14(2):372–386.

5. Sohm JA, et al. (2016) Co-occurring *Synechococcus* ecotypes occupy four major oceanic regimes defined by temperature, macronutrients and iron. *ISME J* 10(2):333–345.

6. Farrant GK, et al. (2016) Delineating ecologically significant taxonomic units from global patterns of marine picocyanobacteria. *Proc Natl Acad Sci* USA 113(24):E3365–E3374.

7. Six C, et al. (2007) Diversity and evolution of phycobilisomes in marine *Synechococcus* spp.: a comparative genomics study. *Genome Biol* 8(12):R259.

8. Alberte RS, Wood AM, Kursar TA, Guillard RRL (1984) Novel phycoerythrins in marine *Synechococcus* spp. *Plant Physiol* 75(3):732–739.

9. Ong LJ, Glazer AN (1991) Phycoerythrins of marine unicellular cyanobacteria. I. Bilin types and locations and energy transfer pathways in *Synechococcus* spp. phycoerythrins. *J Biol Chem* 266(15):9515–9527.

10. Sidler WA (1994) Phycobilisome and phycobiliprotein structures. *The Molecular Biology of Cyanobacteria*, Advances in Photosynthesis. (Springer, Dordrecht), pp 139–216.

11. Humily F, et al. (2013) A gene island with two possible configurations is involved in chromatic acclimation in marine *Synechococcus*. *PLoS ONE* 8(12):e84459.

12. Palenik B (2001) Chromatic adaptation in marine *Synechococcus* strains. *Appl Environ Microbiol* 67(2):991–994.

13. Everroad C, et al. (2006) Biochemical cases of type IV chromatic adaptation in marine *Synechococcus* spp. *J Bacteriol* 188(9):3345–3356.

14. Shukla A, et al. (2012) Phycoerythrin-specific bilin lyase-isomerase controls blue-green chromatic acclimation in marine *Synechococcus*. *Proc Natl Acad Sci* USA 109(49):20136–20141.

15. Sanfilippo JE, et al. (2016) Self-regulating genomic island encoding tandem regulators confers chromatic acclimation to marine *Synechococcus*. *Proc Natl Acad Sci* USA 113(21):6077–6082.

16. Toledo G, Palenik B, Brahamsha B (1999) Swimming marine *Synechococcus* strains with widely different photosynthetic pigment ratios form a monophyletic group. *Appl Environ Microbiol* 65(12):5247–5251.

17. Humily F, et al. (2014) Development of a targeted metagenomic approach to study a genomic region involved in light harvesting in marine *Synechococcus*. *FEMS Microbiol Ecol* 88(2):231–249.

18. Jiang T, et al. (2016) Temporal and spatial variations of abundance of phycocyanin- and phycoerythrin-rich *Synechococcus* in Pearl River Estuary and adjacent coastal area. *J Ocean Univ China* 15(5):897–904.

19. Olson RJ, Chisholm SW, Zettler ER, Armbrust EV (1990) Pigments, size, and distributions of *Synechococcus* in the North Atlantic and Pacific Oceans. *Limnol Oceanogr* 35(1):45–58.

20. Sherry ND, Wood AM (2001) Phycoerythrin-containing picocyanobacteria in the Arabian Sea in february 1995: diel patterns, spatial variability, and growth rates. *Deep Sea Res Part II* 48(6–7):1263–1283.

21. Lantoine F, Neveux J (1997) Spatial and seasonal variations in abundance and spectral characteristics of phycoerythrins in the tropical northeastern Atlantic Ocean. *Deep Sea Res Part I* 44(2):223–246.

22. Neveux J, Lantoine F, Vaulot D, Marie D, Blanchot J (1999) Phycoerythrins in the southern tropical and equatorial Pacific Ocean: evidence for new cyanobacterial types. *J Geophys Res Oceans* 104(C2):3311–3321.

570 23. Campbell L, et al. (1998) Response of microbial community structure to environmental forcing
571 in the Arabian Sea. *Deep Sea Res Part II* 45(10):2301–2325.

572 24. Wood AM, Lipsen M, Coble P (1999) Fluorescence-based characterization of phycoerythrin-
573 containing cyanobacterial communities in the Arabian Sea during the Northeast and early
574 Southwest Monsoon (1994–1995). *Deep Sea Res Part II* 46(8):1769–1790.

575 25. Yona D, Park MO, Oh SJ, Shin WC (2014) Distribution of *Synechococcus* and its phycoerythrin
576 pigment in relation to environmental factors in the East Sea, Korea. *Ocean Sci J* 49(4):367–382.

577 26. Hoge FE, Wright CW, Kana TM, Swift RN, Yungel JK (1998) Spatial variability of oceanic
578 phycoerythrin spectral types derived from airborne laser-induced fluorescence emissions. *Appl*
579 *Opt* 37(21):4744–4749.

580 27. Wood AM, Phinney DA, Yentsch CS (1998) Water column transparency and the distribution of
581 spectrally distinct forms of phycoerythrin- containing organisms. *Mar Ecol Prog Ser* 162:25–31.

582 28. Campbell L, Iturriaga R (1988) Identification of *Synechococcus* spp. in the Sargasso Sea by
583 immunofluorescence and fluorescence excitation spectroscopy performed on individual cells.
584 *Limnol Oceanogr* 33(5):1196–1201.

585 29. Xia X, et al. (2017) Phylogeography and pigment type diversity of *Synechococcus* cyanobacteria
586 in surface waters of the northwestern pacific ocean. *Environ Microbiol* 19(1):142–158.

587 30. Xia X, Liu H, Choi D, Noh JH (2017) Variation of *Synechococcus* pigment genetic diversity along
588 two turbidity gradients in the China Seas. *Microb Ecol*:1–12.

589 31. Xia X, Guo W, Tan S, Liu H (2017) *Synechococcus* assemblages across the salinity gradient in a
590 salt wedge estuary. *Front Microbiol* 8.32. Liu H, Jing H, Wong THC, Chen B (2014) Co-
591 occurrence of phycocyanin- and phycoerythrin-rich *Synechococcus* in subtropical estuarine and
592 coastal waters of Hong Kong: PE-rich and PC-rich *Synechococcus* in subtropical coastal waters.
593 *Environ Microbiol Rep* 6(1):90–99.

594 33. Chung C-C, Gong G-C, Huang C-Y, Lin J-Y, Lin Y-C (2015) Changes in the *Synechococcus*
595 assemblage composition at the surface of the East China Sea due to flooding of the Changjiang
596 river. *Microb Ecol* 70(3):677–688.

597 34. Haverkamp T, et al. (2008) Diversity and phylogeny of Baltic Sea picocyanobacteria inferred
598 from their ITS and phycobiliprotein operons. *Environ Microbiol* 10(1):174–188.

599 35. Stomp M, et al. (2007) Colourful coexistence of red and green picocyanobacteria in lakes and
600 seas. *Ecol Lett* 10(4):290–298.

601 36. Hunter-Cevera KR, Post AF, Peacock EE, Sosik HM (2016) Diversity of *Synechococcus* at the
602 Martha's Vineyard Coastal Observatory: insights from culture isolations, clone libraries, and
603 flow cytometry. *Microb Ecol* 71(2):276–289.

604 37. Fuller NJ, et al. (2003) Clade-specific 16S ribosomal DNA oligonucleotides reveal the
605 predominance of a single marine *Synechococcus* clade throughout a stratified water column in
606 the red sea. *Appl Environ Microbiol* 69(5):2430–2443.

607 38. Larsson J, et al. (2014) Picocyanobacteria containing a novel pigment gene cluster dominate the
608 brackish water Baltic Sea. *ISME J* 8(9):1892–1903.

609 39. Chen F, et al. (2004) Phylogenetic diversity of *Synechococcus* in the Chesapeake Bay revealed by
610 Ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCO) large subunit gene (rbcL)
611 sequences. *Aquat Microb Ecol* 36(2):153–164.

612 40. Choi DH, Noh JH (2009) Phylogenetic diversity of *Synechococcus* strains isolated from the East
613 China Sea and the East Sea. *FEMS Microbiol Ecol* 69(3):439–448.

614 41. Sunagawa S, et al. (2015) Structure and function of the global ocean microbiome. *Science*
615 348(6237):1261359.

616 42. Logares R, et al. (2014) Metagenomic 16S rDNA Illumina tags are a powerful alternative to
617 amplicon sequencing to explore diversity and structure of microbial communities. *Environ*
618 *Microbiol* 16(9):2659–2671.

619 43. Pearman PB, Guisan A, Broennimann O, Randin CF (2008) Niche dynamics in space and time.
620 *Trends Ecol Evol* 23(3):149–158.

621 44. Paulsen ML, et al. (2016) *Synechococcus* in the Atlantic gateway to the Arctic Ocean. *Front Mar*
622 *Sci* 3:191.

623 45. Haverkamp THA, et al. (2008) Colorful microdiversity of *Synechococcus* strains
624 (picocyanobacteria) isolated from the Baltic Sea. *ISME J* 3(4):397–408.

625 46. Cabello-Yeves PJ, et al. (2017) Novel *Synechococcus* genomes reconstructed from freshwater
626 reservoirs. *Front Microbiol* 8:1151.

627 47. Mahmoud RM, et al. (2017) Adaptation to blue light in marine *Synechococcus* requires MpeU,
628 an enzyme with similarity to phycoerythrobilin lyase isomerases. *Front Microbiol* 8:243.

629 48. Veldhuis MJW, Kraay GW (1993) Cell abundance and fluorescence of picoplankton in relation to
630 growth irradiance and nitrogen availability in the red sea. *Neth J Sea Res* 31(2):135–145.

631 49. Katano T, Nakano S (2006) Growth rates of *Synechococcus* types with different phycoerythrin
632 composition estimated by dual-laser flow cytometry in relationship to the light environment in
633 the Uwa Sea. *J Sea Res* 55(3):182–190.

634 50. Ahlgren NA, Rocap G (2006) Culture isolation and culture-independent clone libraries reveal
635 new marine *Synechococcus* ecotypes with distinctive light and N physiologies. *Appl Environ*
636 *Microbiol* 72(11):7193–7204.

637 51. Bemal S, Anil AC (2016) Genetic and ecophysiological traits of *Synechococcus* strains isolated
638 from coastal and open ocean waters of the Arabian Sea. *FEMS Microbiol Ecol* 92(11).

639 52. Everroad RC, Wood AM (2012) Phycoerythrin evolution and diversification of spectral
640 phenotype in marine *Synechococcus* and related picocyanobacteria. *Mol Phylogenet Evol*
641 64(3):381–392.

642 53. Morel A, et al. (2007) Optical properties of the “clearest” natural waters. *Limnol Oceanogr*
643 52(1):217–229.

644 54. Cubillos-Ruiz A, Berta-Thompson JW, Becker JW, van der Donk WA, Chisholm SW (2017)
645 Evolutionary radiation of lanthipeptides in marine cyanobacteria. *Proc Natl Acad Sci* USA
646 114(27):E5424–E5433.

647 55. Martiny AC, Huang Y, Li W (2009) Occurrence of phosphate acquisition genes in
648 *Prochlorococcus* cells from different ocean regions. *Environ Microbiol* 11(6):1340–1347.

649 56. Martiny AC, Coleman ML, Chisholm SW (2006) Phosphate acquisition genes in *Prochlorococcus*
650 ecotypes: evidence for genome-wide adaptation. *Proc Natl Acad Sci* USA 103(33):12552–12557.

651 57. Martiny AC, Kathuria S, Berube PM (2009) Widespread metabolic potential for nitrite and
652 nitrate assimilation among *Prochlorococcus* ecotypes. *Proc Natl Acad Sci* USA 106(26):10787–
653 10792.

58. Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* 30(4):772–780.

59. Guindon S, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59(3):307–321.

60. Huerta-Cepas J, Serra F, Bork P (2016) ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol* 33(6):1635–1638.

61. Camacho C, et al. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

62. Matsen FA, Kodner RB, Armbrust EV (2010) pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 11:538.

63. R Core Team (2014) *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, Austria) Available at: http://www.R-project.org/.

64. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2017) *cluster: cluster analysis basics and extensions*.

65. Venables WN, Ripley BD (2002) *Modern applied statistics with S* (Springer, New York). Fourth edition. Available at: http://www.stats.ox.ac.uk/pub/MASS4.

66. Harrell FE (2016) *Hmisc: Harrell Miscellaneous* Available at: http://CRAN.R-project.org/package=Hmisc.

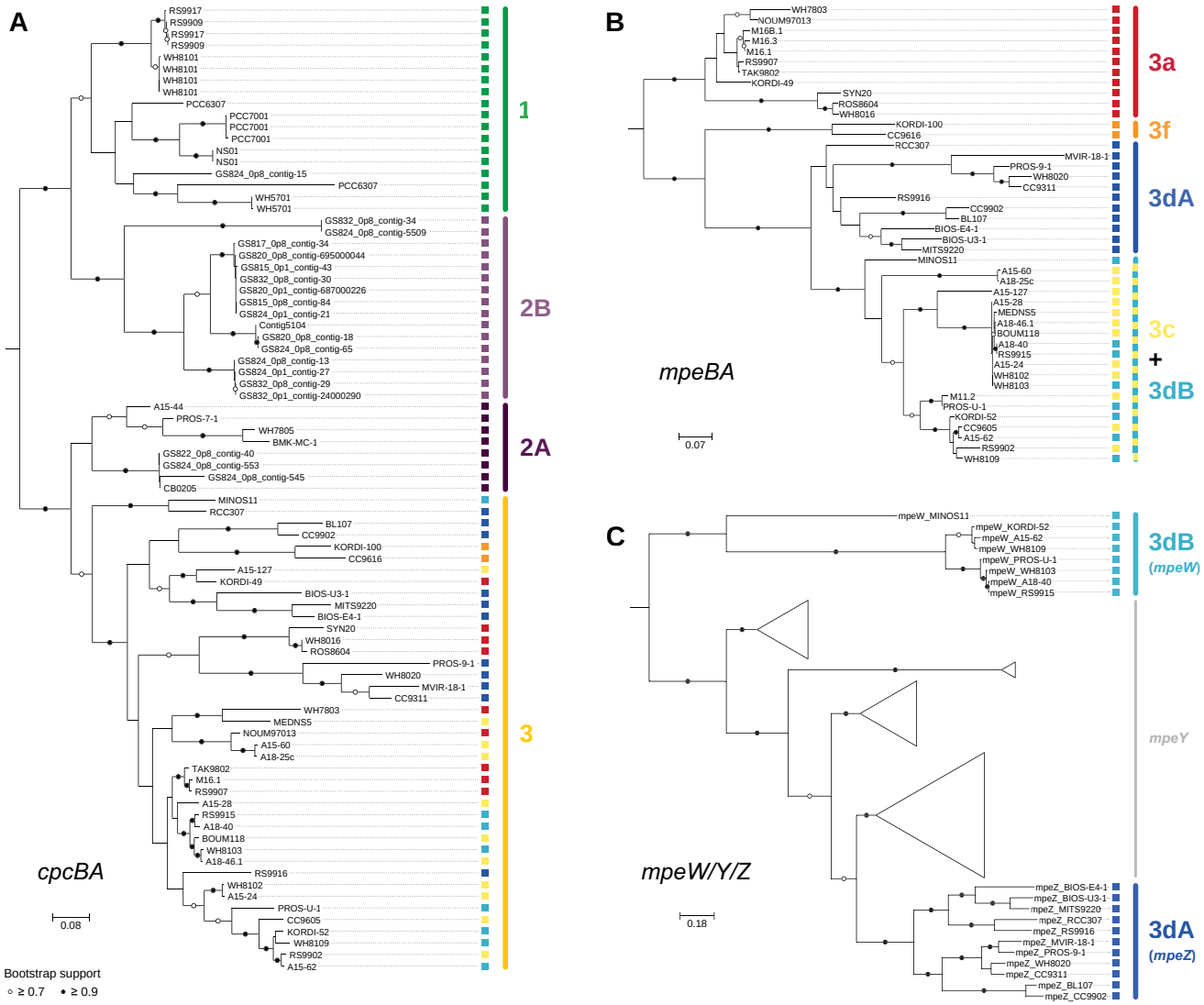
Legends of Figures

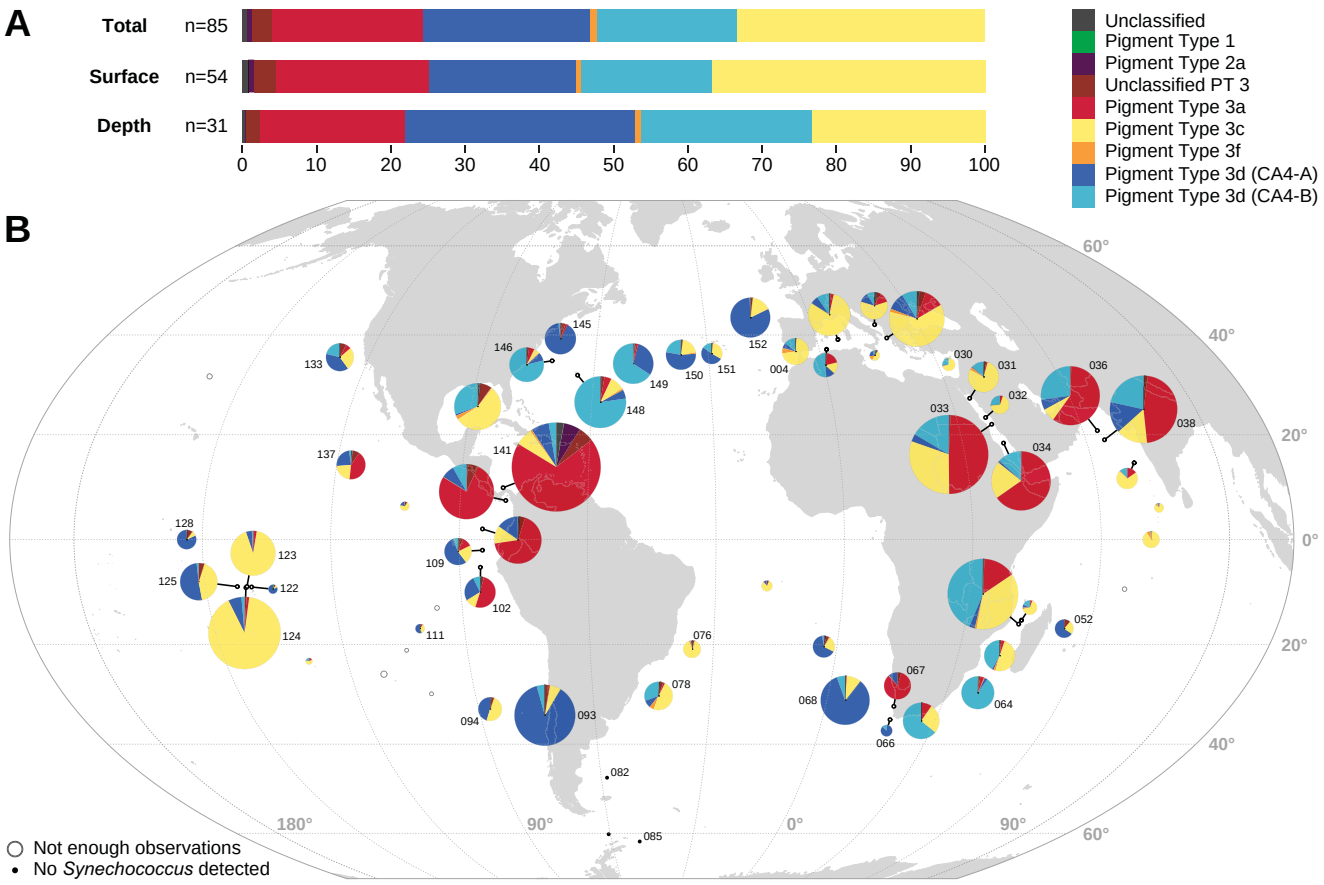
Fig. 1: Maximum likelihood phylogenetic trees of (*A*) *cpcBA* operon, (*B*) *mpeBA* operon and (*C*) the *mpeW/Y/Z* gene family. The *cpcBA* tree includes both strains with characterized pigment type (PT) and environmental sequences (prefixed with GS) assembled from metagenomes of the Baltic Sea (38). Circles at nodes indicate bootstrap support (black: > 90 %; white: > 70 %). Note that for PT 2B clade, only environmental sequences are available. The PT associated with each sequence is indicated as a colored square. The scale bar represents the number of substitutions per nucleotide position.

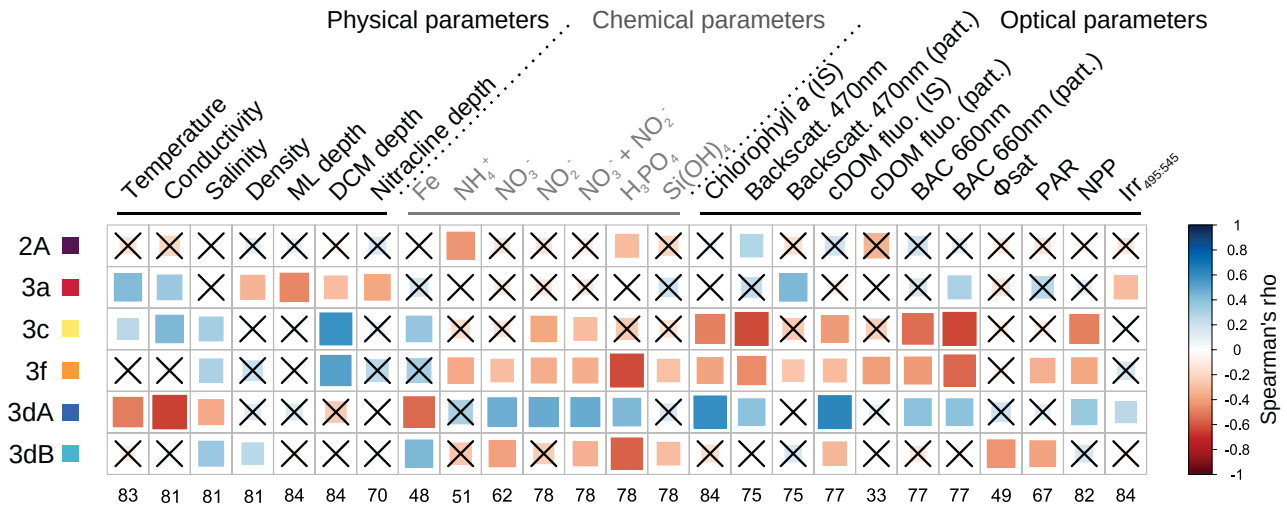
Fig. 2: Distribution of *Synechococcus* pigment types (PT). (*A*) Relative abundance of each PT in the whole dataset (Total), in surface and at the DCM (Deep Chlorophyll Maximum). (*B*) Map showing the global distribution of all *Synechococcus* PTs in surface waters along the *Tara* Oceans transect. Diameters of pies are proportional to the number of *cpcBA* reads normalized by the sequencing effort. Stations with less than 30 *cpcBA* or *mpeBA* reads are indicated by open circles and those with no *cpcBA* reads by black dots. Numbers next to pies correspond to *Tara* Oceans stations.

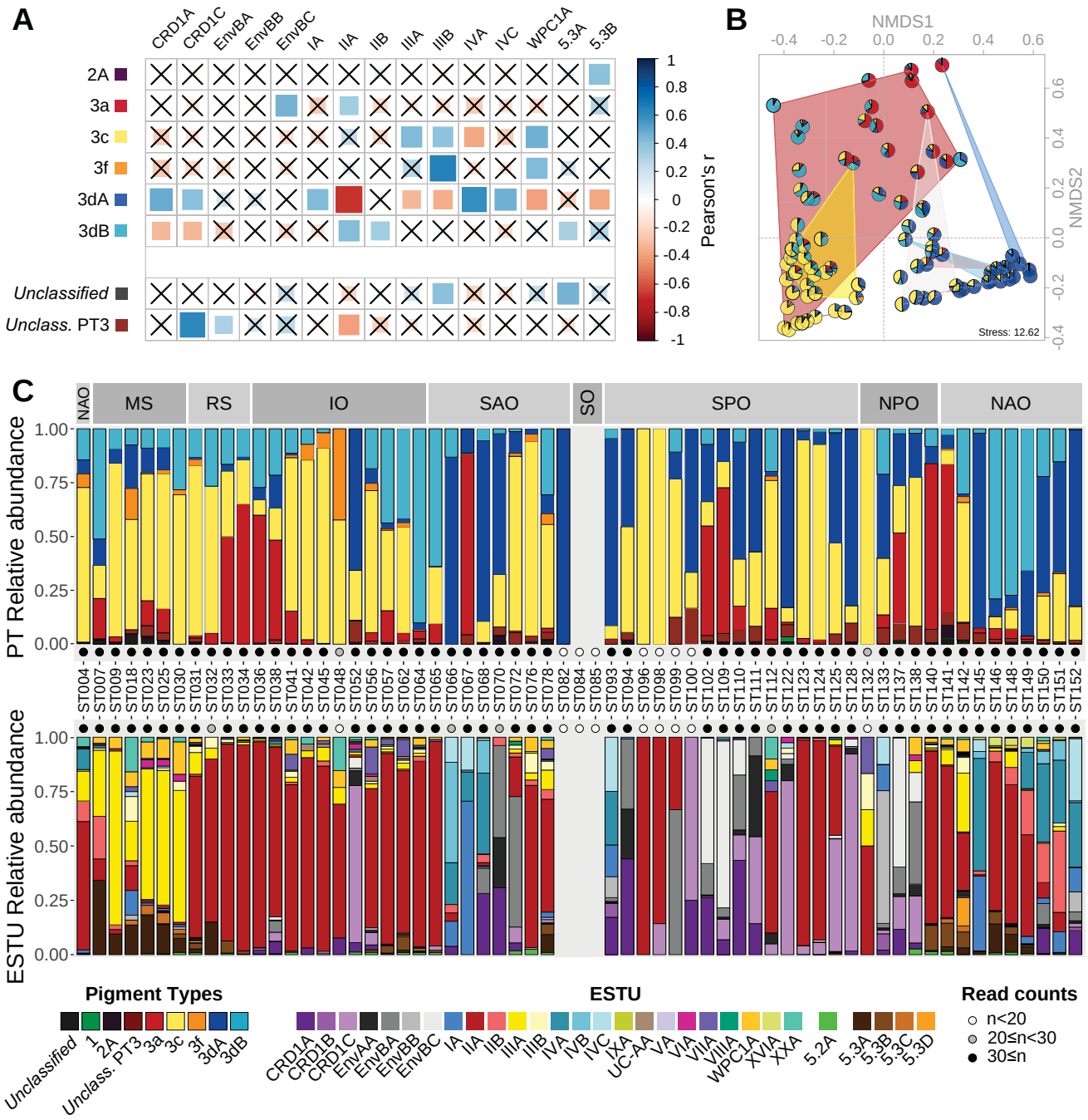
Fig. 3: Correlation analysis between *Synechococcus* pigment types (PT) and environmental parameters measured along the *Tara* Oceans transect for all sampled depths. The scale shows the degree of correlation (blue) or anti-correlation (red) between two variables. Non-significant correlations (adjusted *P* value > 0.05) are indicated by crosses. Number of observations for each environmental parameter is indicated at the bottom. Abbreviations: MLD, mixed layer depth; DCM, deep chlorophyll maximum; IS, *in situ*; Backscatt., backscattering; part., particulate; cDOM fluo, colored dissolved organic matter fluorescence; BAC, beam attenuation coefficient; Φsat, satellite-based non-photochemical quenching (NPQ)-corrected quantum yield of fluorescence (proxy for iron limitation; 6); PAR, photosynthetically active radiation; NPP, net primary production; $Irr_{495:545}$, ratio of downwelling irradiance at 495 nm and 545 nm.

Fig. 4: Relationship between *Synechococcus* pigment types (PT) and Ecologically Significant Taxonomic Units (ESTUs, as defined in 6). (*A*) Correlation analysis between *Synechococcus* PTs and the most abundant ESTUs (>1% relative abundance) for all sampled depths (the complete dataset is shown in Fig. S5). Non-significant correlations (adjusted *P* value > 0.05) are indicated by crosses. The surface of station TARA_067, identified as an outlier (see Fig. S7), was removed for this analysis. (*B*) NMDS analysis of stations according to Bray–Curtis distance between PT assemblages. Samples that belong to the same ESTU assemblage have been contoured with a background color according to the color code used in (6), namely: red, assemblage 1 dominated by ESTU IIA; yellow, assemblage 2 dominated by ESTU IIIA; dark blue, assemblage 4 dominated by ESTUs IA and IVA/B; pink, assemblage 5 co- dominated by ESTUs IIB and IVA/B; grey, assemblage 6 co-dominated by ESTUs CRD1C and EnvBC; light blue, assemblage 8 co-dominated by ESTUs IVA/B, EnvBB and CRD1A/B. (*C*) PT and ESTU relative abundance at each surface station along the *Tara* Oceans transect. Oceanic provinces are indicated in the top gray panels. NAO, North Atlantic Ocean; MS, Mediterranean Sea; RS, Red Sea; IO, Indian Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean; NPO, North Pacific Ocean.


A
RS9917
RS9909
RS9917
RS9909
WH8101
WH8101
WH8101
WH8101
PCC6307
PCC7001
PCC7001
PCC7001
NS01
NS01
GS824_0p8_contig-15
PCC6307
WH5701
WH5701
1
GS832_0p8_contig-34
GS824_0p8_contig-5509
GS817_0p8_contig-34
GS820_0p8_contig-695000044
GS815_0p1_contig-43
GS832_0p8_contig-30
GS820_0p1_contig-687000226
GS815_0p8_contig-84
GS824_0p1_contig-21
Contig5104
GS820_0p8_contig-18
GS824_0p8_contig-65
GS824_0p8_contig-13
GS824_0p1_contig-27
GS832_0p8_contig-29
GS832_0p1_contig-24000290
2B
A15-44
PROS-7-1
WH7805
BMK-MC-1
GS822_0p8_contig-40
GS824_0p8_contig-553
GS824_0p8_contig-545
CB0205
2A
MINOS11
RCC307
BL107
CC9902
KORDI-100
CC9616
A15-127
KORDI-49
BIOS-U3-1
MITS9220
BIOS-E4-1
SYN20
WH8016
ROS8604
PROS-9-1
WH8020
MVIR-18-1
CC9311
WH7803
MEDNS5
NOUM97013
A15-60
A18-25c
TAK9802
M16.1
RS9907
A15-28
RS9915
A18-40
BOUM118
WH8103
A18-46.1
RS9916
WH8102
A15-24
PROS-U-1
CC9605
KORDI-52
WH8109
RS9902
A15-62
3
cpcBA
0.08
Bootstrap support
◦ ≥ 0.7 • ≥ 0.9
B
WH7803
NOUM97013
M16B.1
M16.3
M16.1
RS9907
TAK9802
KORDI-49
SYN20
ROS8604
WH8016
KORDI-100
CC9616
RCC307
MVIR-18-1
PROS-9-1
WH8020
CC9311
RS9916
CC9902
BL107
BIOS-E4-1
BIOS-U3-1
MITS9220
MINOS11
A15-60
A18-25c
A15-127
A15-28
MEDNS5
A18-46.1
BOUM118
A18-40
RS9915
A15-24
WH8102
WH8103
M11.2
PROS-U-1
KORDI-52
CC9605
A15-62
RS9902
WH8109
3a
3f
3dA
3c
+
3dB
mpeBA
0.07
C
mpeW_MINOS11
mpeW_KORDI-52
mpeW_A15-62
mpeW_WH8109
mpeW_PROS-U-1
mpeW_WH8103
mpeW_A18-40
mpeW_RS9915
3dB
(mpeW)
mpeY
mpeZ_BIOS-E4-1
mpeZ_BIOS-U3-1
mpeZ_MITS9220
mpeZ_RCC307
mpeZ_RS9916
mpeZ_MVIR-18-1
mpeZ_PROS-9-1
mpeZ_WH8020
mpeZ_CC9311
mpeZ_BL107
mpeZ_CC9902
3dA
(mpeZ)
mpeW/Y/Z
0.18


A
Total n=85
Surface n=54
Depth n=31
0 10 20 30 40 50 60 70 80 90 100
Unclassified
Pigment Type 1
Pigment Type 2a
Unclassified PT 3
Pigment Type 3a
Pigment Type 3c
Pigment Type 3f
Pigment Type 3d (CA4-A)
Pigment Type 3d (CA4-B)
B
60°
40°
20°
0°
20°
40°
60°
180°
90°
0°
90°
133 145 146 149 148 150 151 152 004 030 031 036 038 032 033 034 137 141 128 123 125 122 124 109 102 111 076 078 093 094 067 068 066 064 052 082 085
Not enough observations
No Synechococcus detected


Physical parameters
Chemical parameters
Optical parameters
Temperature
Conductivity
Salinity
Density
ML depth
DCM depth
Nitracline depth
Fe
NH_4^+
NO_3^-
NO_2^-
$NO_3^- + NO_2^-$
H_3PO_4
$Si(OH)_4$
Chlorophyll a (IS)
Backscatt. 470nm
Backscatt. 470nm (part.)
cDOM fluo. (IS)
cDOM fluo. (part.)
BAC 660nm
BAC 660nm (part.)
Φsat
PAR
NPP
$Irr_{495:545}$
2A
3a
3c
3f
3dA
3dB
83 81 81 81 84 84 70 48 51 62 78 78 78 78 84 75 75 77 33 77 77 49 67 82 84
Spearman's rho
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


A
CRD1A
CRD1C
EnvBA
EnvBB
EnvBC
IA
IIA
IIB
IIIA
IIIB
IVA
IVC
WPC1A
5.3A
5.3B
2A
3a
3c
3f
3dA
3dB
Unclassified
Unclass. PT3
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
Pearson's r
B
NMDS1
-0.4
-0.2
0.0
0.2
0.4
0.6
NMDS2
Stress: 12.62
C
NAO
MS
RS
IO
SAO
SO
SPO
NPO
NAO
PT Relative abundance
ESTU Relative abundance
1.00
0.75
0.50
0.25
0.00
ST004
ST007
ST009
ST018
ST023
ST025
ST030
ST031
ST032
ST033
ST034
ST036
ST038
ST041
ST042
ST045
ST048
ST052
ST056
ST057
ST062
ST064
ST065
ST066
ST067
ST068
ST070
ST072
ST076
ST078
ST082
ST084
ST085
ST093
ST094
ST096
ST098
ST099
ST100
ST102
ST109
ST110
ST111
ST112
ST122
ST123
ST124
ST125
ST128
ST132
ST133
ST137
ST138
ST140
ST141
ST142
ST145
ST146
ST148
ST149
ST150
ST151
ST152
Pigment Types
Unclassified
1
2A
Unclass. PT3
3a
3c
3f
3dA
3dB
ESTU
CRD1A
CRD1B
CRD1C
EnvAA
EnvBA
EnvBB
EnvBC
IA
IIA
IIB
IIIA
IIIB
IVA
IVB
IVC
IXA
UC-AA
VA
VIA
VIIA
VIIIA
WPC1A
XVIA
XXA
5.2A
5.3A
5.3B
5.3C
5.3D
Read counts
n<20
20≤n<30
30≤n

Supplementary materials and methods

Modeling of the blue to green irradiance ratio ($Irr_{495:545}$) at *Tara* Oceans stations.

We used the clear sky surface irradiance model of Frouin and McPherson in Fortran and translated to Matlab by Werdell (see Frouin et al., 1989 and Tanre et al., 1979 for the analytical formula used) using the date, latitude and longitude of each station, assuming sunny sky and at noon.

The spectral light distribution averaged over the mixed layer was computed from:

$$< Ir(\lambda) >= \frac{\int_0^{MLD} E(\lambda,0^-)e^{-k(\lambda,chl)z}dz}{MLD} = \frac{I(\lambda,0^-)}{MLD\ k(\lambda,chl)}\{1 - e^{-k(chl)MLD}\}$$

where:
- *chl* denotes the average chlorophyll value in the mixed layer. [*chl*] was based on a fluorometer that was calibrated against HPLC data and corrected for non-photochemical quenching,
- MLD is the mixed layer depth that was computed based on a temperature threshold criterion
- $k(\lambda, chl)$ is the diffuse attenuation coefficient at wavelength λ (495 or 545 using a 10 nm bandwidth). This parameter was computed using Morel and Maritorena (2001)'s equation:

$$k(\lambda, chl) = k_w(\lambda) + \chi(\lambda)[chl]^{e(\lambda)}$$

k_w, χ and e are provided in Table 2 of Morel and Maritorena (2001) and have the following values for the wavelengths of interest:

Wavelength [nm]	$k_w(\lambda)$ [m^{-1}]	$\chi(\lambda)$	$e(\lambda)$
495	0.01885	0.06907	0.68947
545	0.05212	0.04253	0.65591

If the sampling depth was below the MLD, the irradiance was computed as follows:

$$Ir(\lambda, sampling\ depth) = (\lambda, 0^-)e^{-k(\lambda,chl)sampling\ depth}.$$

The ratio was then computed as $Irr_{495:545}$.

References:

Frouin, R., D. W. Ligner, and C. Gautier, 1989: A Simple analytical formula to compute clear sky total and photosynthetically available solar irradiance CC at the ocean surface. J. Geophys. Res., 94, 9731-9742.

Morel, A. and S. Maritorena, 2001: Bio-optical properties of oceanic waters: A reappraisal. J. Geophys. Res., 106, 7163–7180.

Tanre, D., M. Herman, P.-Y. Deschamps, and A. De Leffe, 1979: Atmospheric modeling for Space measurements of ground reflectances, including bi-directional properties. Appl. Optics, 18, 21,3587-21,3597.

Legends to supplementary figures

Figure S1: Biochemical composition and biooptical properties of phycobilisomes (PBS) of the main *Synechococcus* pigment types (PT). *(A)* Models of PBS structure, highlighting the conserved core and variable rods of increasing complexity from PT1 to PT3 (Redrawn after Six *et al.*, 2007). *(B)* Whole cell absorption spectra of the different PTs (Reproduced after Six *et al.*, 2007). Chromophores responsible of each absorption peaks are indicated. *(C)* Whole cell fluorescence excitation spectra with emission at 680 nm. Note that for chromatic acclimaters (PT 3d), the PBS structure is similar to other PT 3 but that the excitation ratio at 495 nm and 545 nm ($Ex_{495:545}$) varies from 0.6 in green light to 1.6 in blue light (not shown).

Figure S2: Evaluation of the assignment pipeline and the resolution power of the different markers used in this study. Simulated reads were generated from the reference dataset and assigned using a custom-designed pipeline (see materials and methods). *(A, C, E)* Evaluation of different sets of parameters tested for read assignment for the different markers: *cpcBA (A)*, *mpeBA (C)* and *mpeW (E)*. 100 (yellow), 125 (pink) and 150 bp (dark red) long reads were simulated. For each read, pplacer returns a list of possible positions in the tree, each associated with a likelihood weight. From these placements, we considered only those that reached a summed likelihood weight of either 90% (circle), 95% (square) or 100% (triangle). The assignment was then performed based on the phenotype of either the nearest node (solid symbol) in the tree or the descending (child) node (empty symbol). *(B, D, F)* Evaluation the resolution power along *cpcBA (B)*, *mpeBA (D)* or *mpeW (F)* for 150 bp simulated reads assigned using the parameters selected for *Tara* Oceans metagenomic read assignment (i.e., nearest node assignment and summed weight of 95%). Note that *Tara* Oceans reads had a mean length of 164 bp.

Figure S3: Correlations between the number of reads recruited using the main markers used in this study. *(A)* Correlation between *petB* (vertical phylogeny) and *cpcBA* counts used to discriminate pigment types (PT) 1, 2 and 3. *(B)* Correlation between PT 3 counts using *cpcBA* and total *mpeBA* counts. Note that *mpeBA* is a PT3 specific marker and is used to discriminate PTs 3a, 3dA, 3f and 3c+3dB. *(C)* Correlation between PT 3dB counts using *fciAB*, a PT 3dB- and 3dA-specific marker and total *mpeW* counts, a PT 3dB-specific marker.

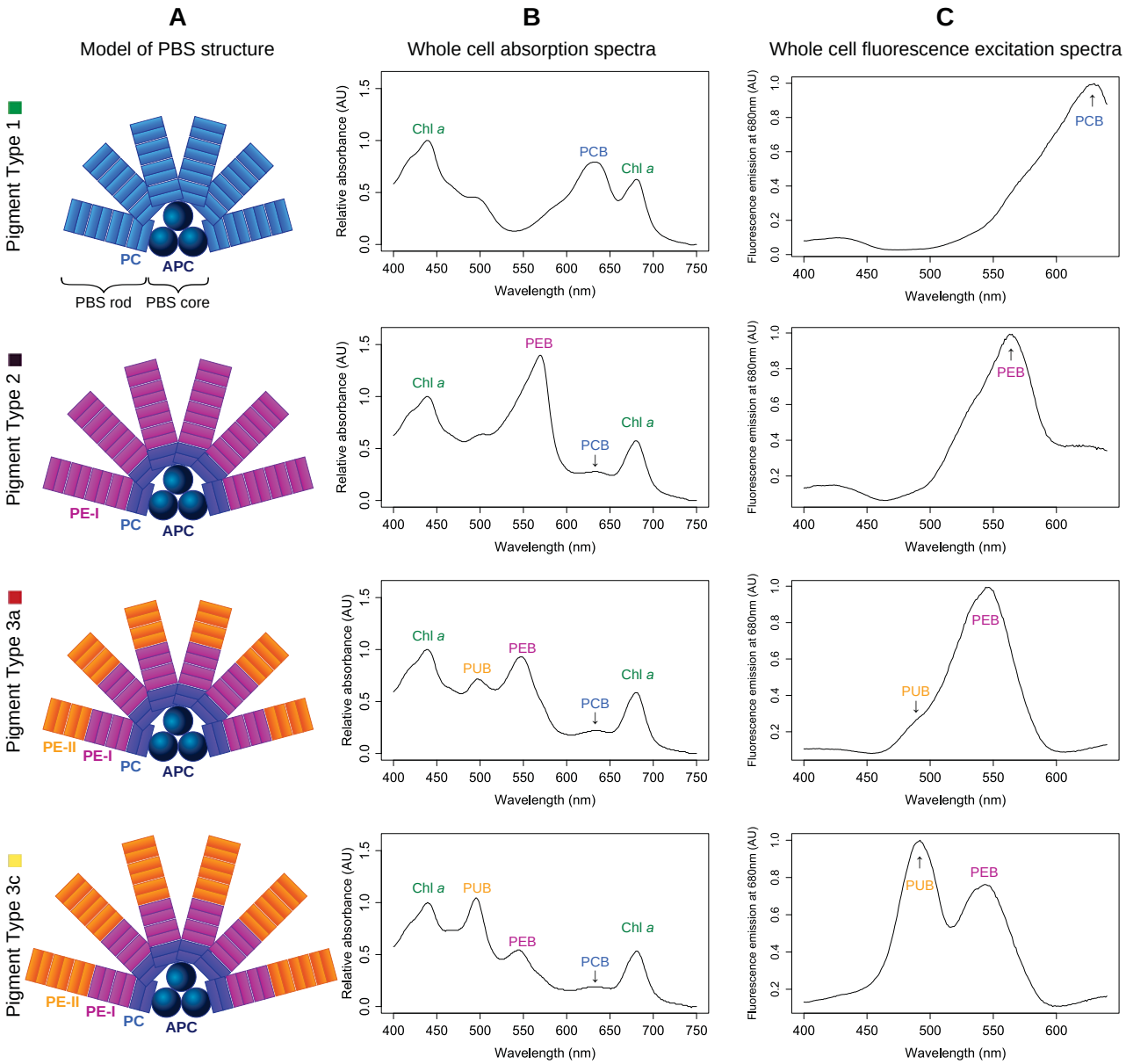
Figure S4: Distribution of *Synechococcus* pigment types (PTs) at depth (Deep Chlorophyll Maximum). *(A)* Map showing the global distribution of all *Synechococcus* PTs at depth along the *Tara* Oceans transect. Diameters of pies are proportional to the number of *cpcBA* reads normalized by the sequencing effort. Stations with less than 30 *cpcBA* or *mpeBA* reads are indicated by open circles and those with no *cpcBA* reads by black dots. Numbers next to pies correspond to *Tara* Oceans stations. *(B)* PTs and ESTU relative abundance at depth for sampling station along the *Tara* Oceans transect. Oceanic provinces are indicated in the top gray panels. NAO, North Atlantic Ocean; MS, Mediterranean Sea; RS, Red Sea; IO, Indian Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean; NPO, North Pacific Ocean.

Figure S5: Same as Fig. 4*A* but for all ESTUs. Unclass., unclassified.

Figure S6: Focus on pigment type (PT) 3dA natural mutants, exhibiting an altered gene content with regard to typical PT 3dA. *(A-J)* Correlation between the number of reads assigned as PT 3dA using different markers (all present in single gene copy in typical 3dA). Each circle corresponds to a *Tara* Oceans station and depth. Orange circles: stations with at least 20 *mpeBA* reads assigned to PT 3dA and at least twice more 3dA counted with *mpeBA* than with *fciAB*, corresponding to the surface sample of stations TARA_070, TARA_110 and TARA_137 and the

DCM of stations TARA_038, TARA_058 and TARA_110. Red circles: same but with more than 10-fold 3dA counted with *mpeBA* than *fciAB*, corresponding to the surface sample of stations TARA_052, TARA_094, TARA_111 and TARA_122 to TARA_128, and DCM of stations TARA_052, TARA_100, TARA_111 and TARA_128. Green circle: surface of station TARA_067. *(K)* CA4-A genomic island and fragment of the phycobilisome (PBS) genomic region for a typical, CA4-able 3dA strain (strain BL107), and 3 CA4-deficient strains, which are stuck either in blue light phenotype (similar to strain BIOS-E4-1), or green light phenotype (as strains MVIR-18-1 and WH8016). Note that KORDI-49 and WH8016 strains have identical PBS gene complement and genomic arrangement. The complete PBS genomic region of the BL107 strain can be found in Six *et al.*, 2007. Note that for readability, surface of station TARA_093 has been omitted since it has the highest normalized counts (2.7-3.2) for all markers and exhibited a good agreement between markers (ratio close to 1:1).

Figure S7: Correlation between the proportion of clades I, IV and CRD1, as assessed with *petB*, and the proportion of pigment type 3dA, as assessed with *mpeBA*, at each station.



Phycobiliproteins

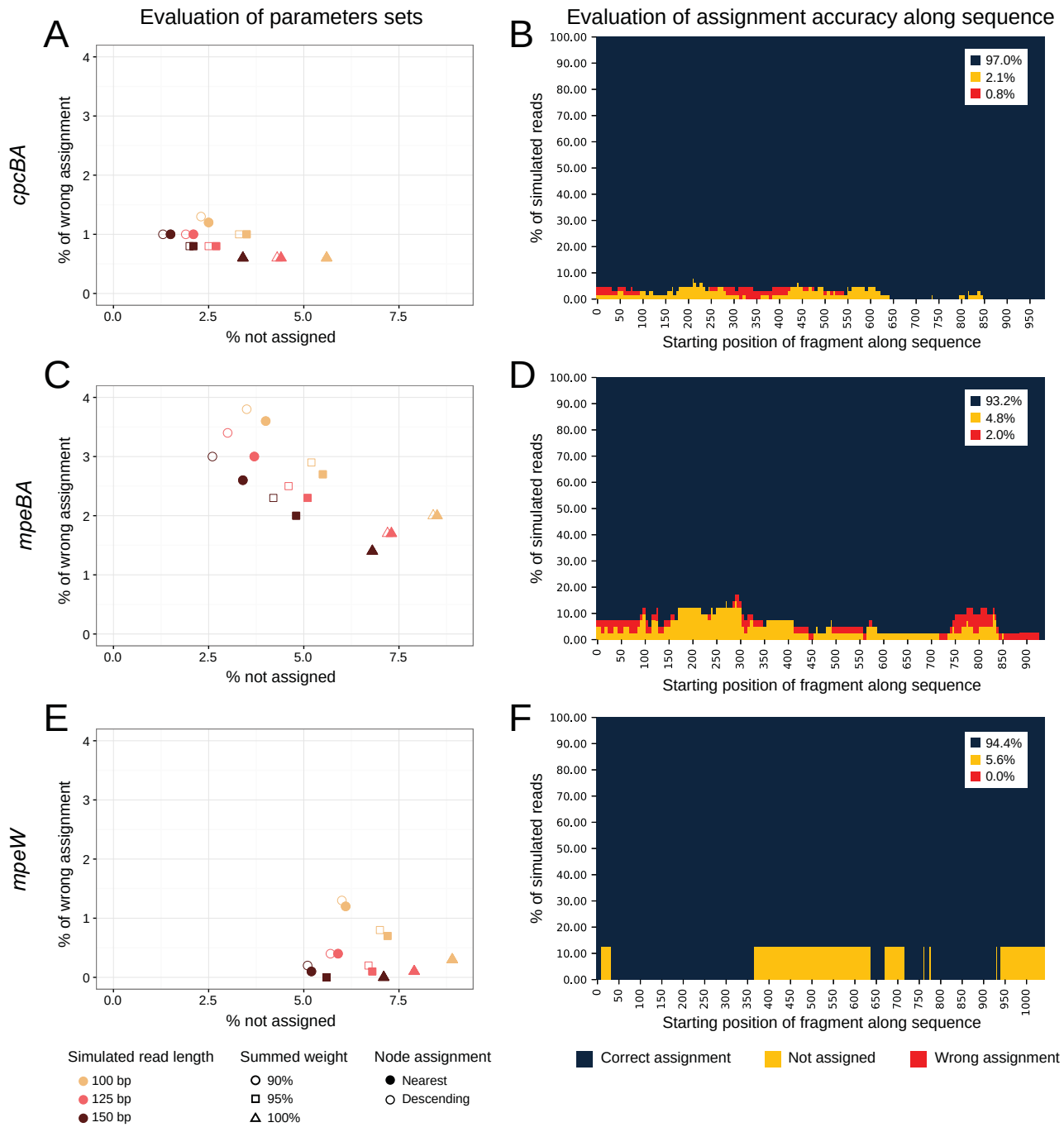
APC	Allophycocyanin
PC	Phycocyanin
PE-I	Phycoerythrin-I
PE-II	Phycoerythrin-II

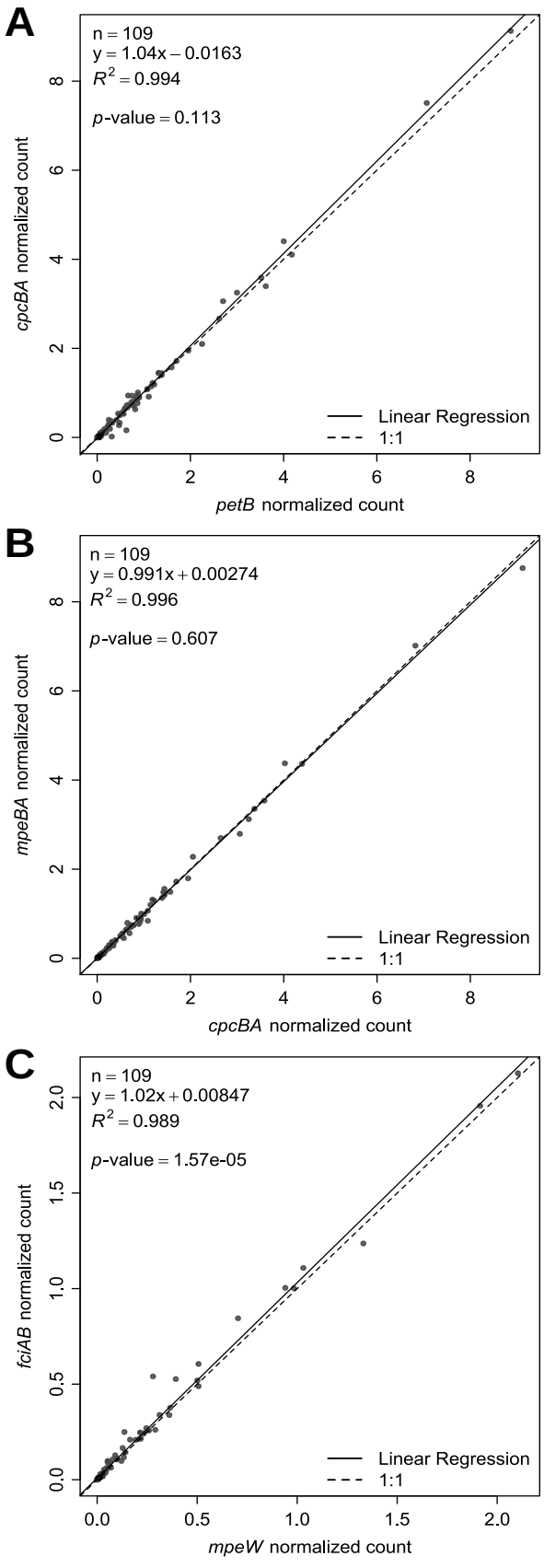
Phycobilins

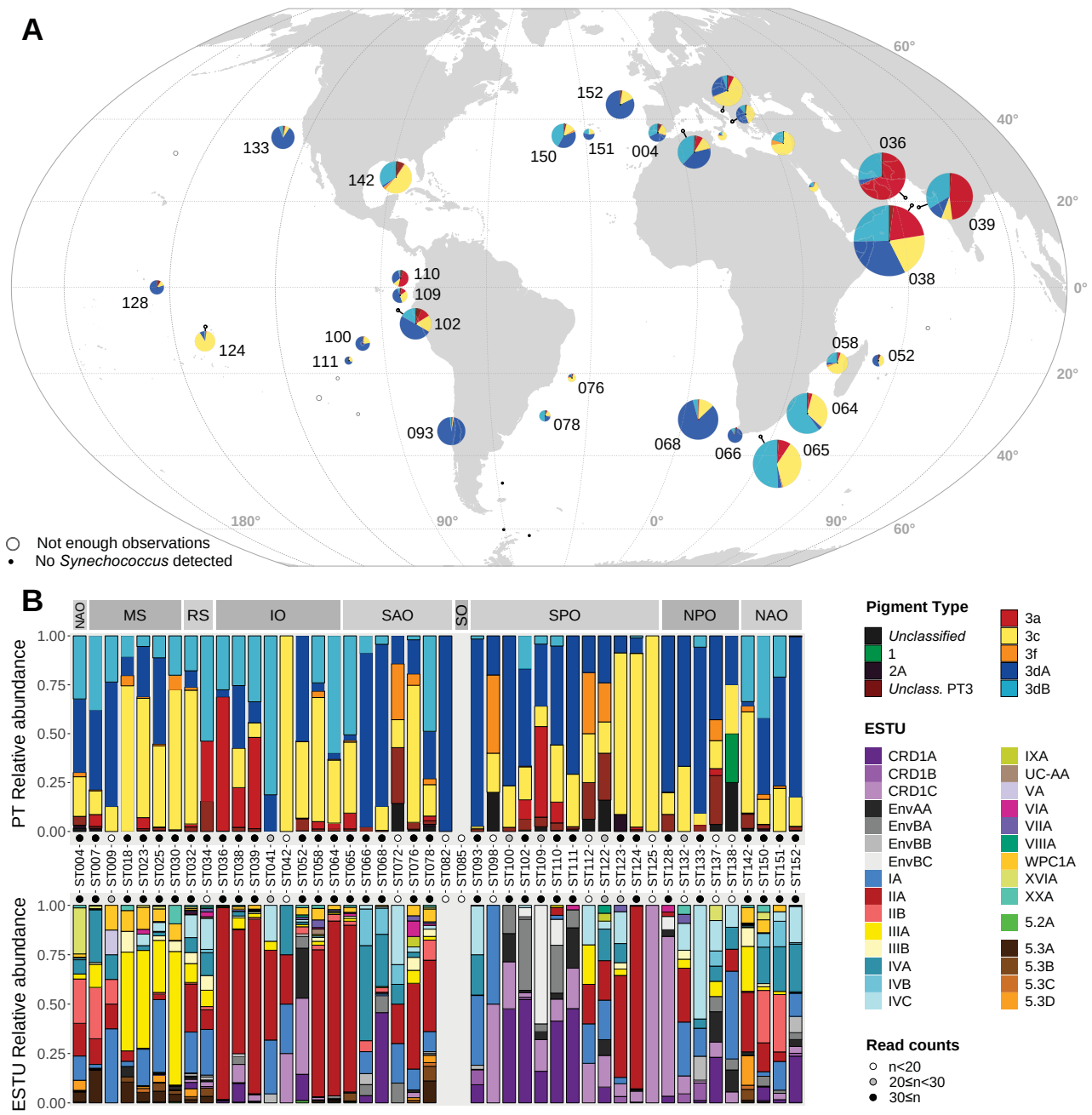
PCB	Phycocyanobilin (A_{max} = 620 nm)
PEB	Phycoerythrobilin (A_{max} = 545-560 nm)
PUB	Phycourobilin (A_{max} = 495 nm)

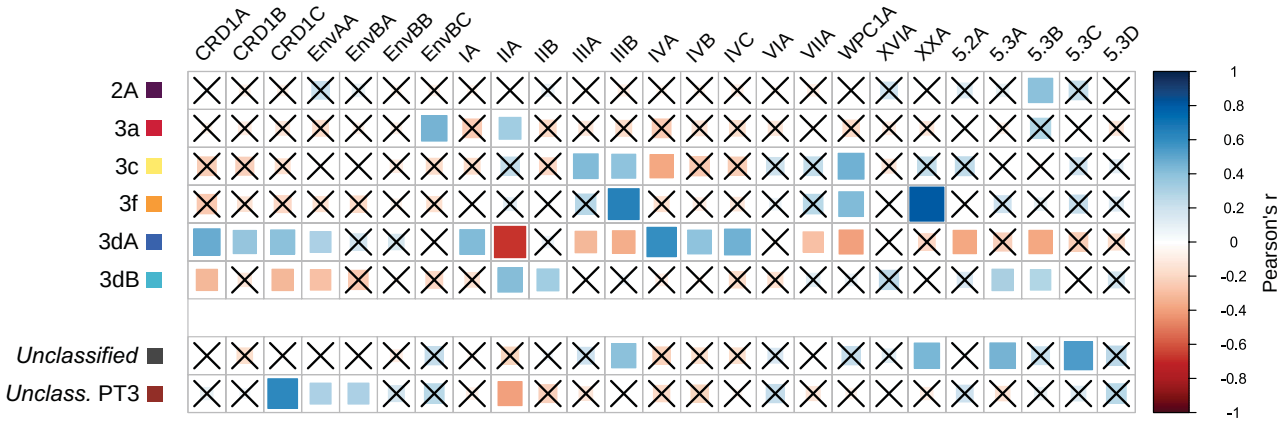
Other chromophore

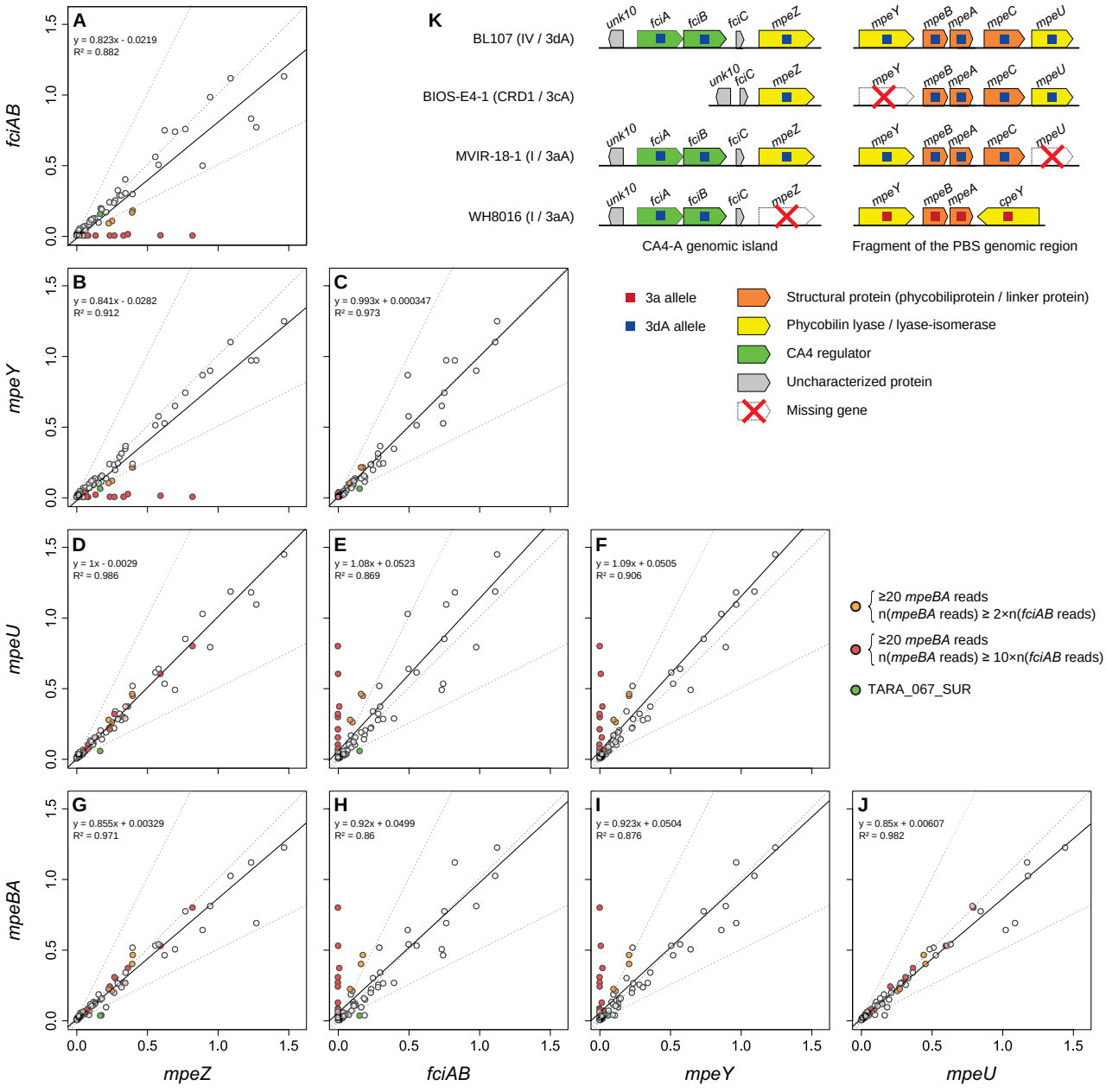
Chl *a*	Chlorophyll *a*

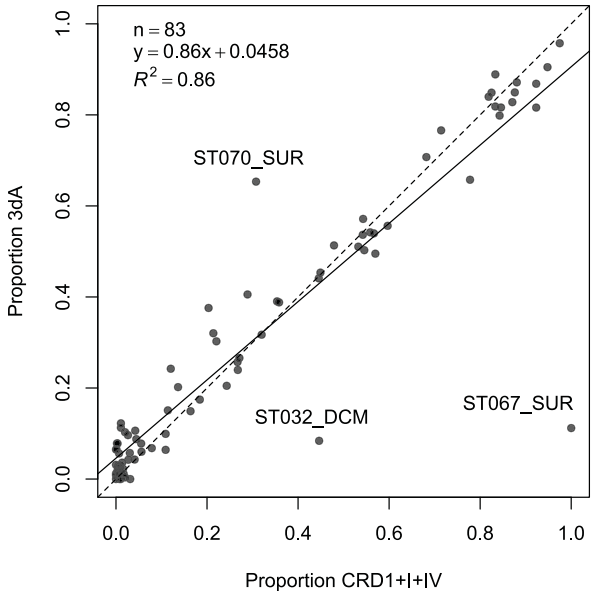

Evaluation of parameters sets
Evaluation of assignment accuracy along sequence
A
B
C
D
E
F
cpcBA
mpeBA
mpeW
% of wrong assignment
% not assigned
% of simulated reads
Starting position of fragment along sequence
97.0%
2.1%
0.8%
93.2%
4.8%
2.0%
94.4%
5.6%
0.0%
Simulated read length
100 bp
125 bp
150 bp
Summed weight
90%
95%
100%
Node assignment
Nearest
Descending
Correct assignment
Not assigned
Wrong assignment


A
n = 109
y = 1.04x – 0.0163
R^2 = 0.994
p-value = 0.113
Linear Regression
1:1
cpcBA normalized count
petB normalized count
B
n = 109
y = 0.991x + 0.00274
R^2 = 0.996
p-value = 0.607
Linear Regression
1:1
mpeBA normalized count
cpcBA normalized count
C
n = 109
y = 1.02x + 0.00847
R^2 = 0.989
p-value = 1.57e-05
Linear Regression
1:1
fciAB normalized count
mpeW normalized count


A
60°
40°
20°
0°
20°
40°
60°
180°
90°
0°
90°
152
133
150
151
004
036
142
039
038
110
109
128
102
124
100
111
058
052
076
064
093
078
068
066
065
Not enough observations
No Synechococcus detected
B
NAO
MS
RS
IO
SAO
SO
SPO
NPO
NAO
PT Relative abundance
ESTU Relative abundance
1.00
0.75
0.50
0.25
0.00
ST004 ST007 ST009 ST018 ST023 ST025 ST030 ST032 ST034 ST036 ST038 ST039 ST041 ST042 ST052 ST058 ST064 ST065 ST066 ST068 ST072 ST076 ST078 ST082 ST085 ST093 ST098 ST100 ST102 ST109 ST110 ST111 ST112 ST122 ST123 ST124 ST125 ST128 ST132 ST133 ST137 ST138 ST142 ST150 ST151 ST152
Pigment Type
Unclassified
1
2A
Unclass. PT3
3a
3c
3f
3dA
3dB
ESTU
CRD1A
CRD1B
CRD1C
EnvAA
EnvBA
EnvBB
EnvBC
IA
IIA
IIB
IIIA
IIIB
IVA
IVB
IVC
IXA
UC-AA
VA
VIA
VIIA
VIIIA
WPC1A
XVIA
XXA
5.2A
5.3A
5.3B
5.3C
5.3D
Read counts
n<20
20≤n<30
30≤n


CRD1A
CRD1B
CRD1C
EnvAA
EnvBA
EnvBB
EnvBC
IA
IIA
IIB
IIIA
IIIB
IVA
IVB
IVC
VIA
VIIA
WPC1A
XVIA
XXA
5.2A
5.3A
5.3B
5.3C
5.3D
2A
3a
3c
3f
3dA
3dB
Unclassified
Unclass. PT3
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
Pearson's r


A
y = 0.823x - 0.0219
R² = 0.882
B
y = 0.841x - 0.0282
R² = 0.912
C
y = 0.993x + 0.000347
R² = 0.973
D
y = 1x - 0.0029
R² = 0.986
E
y = 1.08x + 0.0523
R² = 0.869
F
y = 1.09x + 0.0505
R² = 0.906
G
y = 0.855x + 0.00329
R² = 0.971
H
y = 0.92x + 0.0499
R² = 0.86
I
y = 0.923x + 0.0504
R² = 0.876
J
y = 0.85x + 0.00607
R² = 0.982
fciAB
mpeY
mpeU
mpeBA
mpeZ
fciAB
mpeY
mpeU
K
BL107 (IV / 3dA)
BIOS-E4-1 (CRD1 / 3cA)
MVIR-18-1 (I / 3aA)
WH8016 (I / 3aA)
unk10
fciA
fciB
fciC
mpeZ
mpeY
mpeB
mpeA
mpeC
mpeU
cpeY
CA4-A genomic island
Fragment of the PBS genomic region
3a allele
3dA allele
Structural protein (phycobiliprotein / linker protein)
Phycobilin lyase / lyase-isomerase
CA4 regulator
Uncharacterized protein
Missing gene
≥20 mpeBA reads
n(mpeBA reads) ≥ 2×n(fciAB reads)
≥20 mpeBA reads
n(mpeBA reads) ≥ 10×n(fciAB reads)
TARA_067_SUR


n = 83
y = 0.86x + 0.0458
R^2 = 0.86
ST070_SUR
ST032_DCM
ST067_SUR
Proportion 3dA
Proportion CRD1+I+IV
0.0
0.2
0.4
0.6
0.8
1.0